Logistics Management Institute

A Physics-Based Alternative to Cost-Per-Flying-Hour Models of Aircraft Consumption Costs

AF909T1

August 2000

John M. Wallace
Scott A. Houser
David A. Lee

LMI®

REPORT DOCUMENTATION PAGE

Form Approved
OPM No.0704-0188

1. AGENCY USE ONLY *(Leave Blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	Aug 00	Final

4. TITLE AND SUBTITLE

A Physics-Based Alternative to Cost-Per-Flying-Hour Models of Aircraft Consumption Costs

5. FUNDING NUMBERS

C -GS-23F-9737H

PE 0902198D

6. AUTHOR(S)

John M. Wallace, Scott A. Houser, David A. Lee

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Logistics Management Institute
2000 Corporate Ridge
McLean, VA 22102-7805

8. PERFORMING ORGANIZATION REPORT NUMBER

LMI- AF909T1

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Cost Analysis Agency
SAF/SMC
1130 Air Force Pentagon
Washington DC 20330-1130

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

A: Approved for public release; distribution is unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

Data from Operation Desert Storm shows that current aircraft consumption models based upon a historical Cost Per Flying Hour (CPFH) are grossly inaccurate in predicting consumption cost during wartime surges. This report proposes an alternative to CPFH-based models that considers other parameters that drive consumption costs, namely time on the ground, sorties, and landings. We tested the model using C-5B transport data from Operation Desert Storm, and verified it using C-17 transport, KC-10 all-purpose tanker, and F-16 fighter data from recent years, which includes Kosovo operations. Our investigation shows that this physics-based model is at least as accurate as the CPFH-based model in the general case, is far more accurate during wartime surges, and is generally more robust than the CPFH-based model. We also found that the physics-based model can be feasibly implemented and used by United States Air Force analysts.

14. SUBJECT TERMS

Cost Per Flying Hour, Aircraft consumption modeling, Operation Desert Storm, Kosovo air campaigns

15. NUMBER OF PAGES

58

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL

A Physics-Based Alternative to Cost-Per-Flying-Hour Models of Aircraft Consumption Costs

AF909T1

August 2000

John M. Wallace
Scott A. Houser
David A. Lee

LOGISTICS MANAGEMENT INSTITUTE
2000 CORPORATE RIDGE
MCLEAN, VIRGINIA 22102-7805

A Physics-Based Alternative to Cost-Per-Flying-Hour Models of Aircraft Consumption Costs

AF909T1/AUGUST 2000

Executive Summary

Operation Desert Storm (ODS) illustrated problems associated with proportional models—based on computing a cost per flying hour (CPFH) for each fleet of aircraft—to estimate wartime resource consumption. During ODS, these models overpredicted the amount of materials consumed by more than 200 percent. Faced with the requirement to develop realistic supplemental budgets for future contingencies (such as Kosovo), the Assistant Secretary of the Air Force for Cost and Economics asked LMI to study alternatives to the CPFH methodology.

We used C–5B data from ODS to formulate a physics-based model that considers three separate failure modes:

- Dormant failures that relate to failures induced while an airplane is on the ground (measured in terms of ground time)
- Cycle-induced failures from stress-inducing events that are not related to how long the aircraft operates (measured in terms of sorties and takeoffs and landings)
- Failures associated with operating time (measured in flying hours).

We verified the model with Kosovo-era data for the C–17 transport, the KC–10 all-purpose tanker, and the F–16C fighter. (The C–5B fleet did not fly a significant wartime surge during Kosovo; therefore, we did not use it to verify our model.)

Our analysis showed that the physics-based model significantly outperformed the proportional CPFH model during surges; it performed at least as well as the proportional model in all cases. Equally important, the physics-based model is more robust than the CPFH model: Its accuracy does not degrade radically after a major flight program change, whereas a CPFH model typically does. The physics-based model uses data from existing data sources and requires no specialized software (other than a spreadsheet program such as Excel).

Our investigation made an important discovery that will improve the accuracy of physics-based and proportional models alike. Three of the four aircraft fleets that we studied have a *removals bathtub*: In the first few years of service, a fleet exhibits a gradual decrease in removals that cannot be explained by changes in flight characteristics. As the fleet matures, this trend levels out. Beyond a certain age, removals increase. This trend can last for several years. Current models do not take these gradual trends into account, which can cause serious errors in forecasting resource consumption. Our study suggests an approach for incorporating these trends in resource models. We need further study, however, to establish rules for identifying these trends, estimating their magnitude, and determining when they start and stop.

We suggest that the Air Force extend this important research program, focusing on four areas:

- The current physics-based model has just four parameters. We should investigate whether the inclusion of additional parameters, such as accelerations above a certain g-force or the use of afterburners, can improve the model.
- The F–16C data from Aviano Air Base differ significantly from the full-fleet data. Now that we know that the model is valid in general, we can use it to investigate base-to-base variations.
- We used a linear shear transformation to adjust for removals bathtubs. Do other transformations afford better results?
- We should develop software that automates the tasks of obtaining input data, calibrating the model, and investigating the results. LMI has successfully developed several such applications to support other analyses.

Contents

FIGURES

TABLE

Chapter 1
Introduction

Current maintenance models that are based on aircraft flying hours do not work when flying behavior changes significantly. These proportional models predict the maintenance needs of a fleet of aircraft on the basis of a simple scaling method. They do not consider the physics of removal-causing failures; therefore, they cannot accurately predict material consumption during periods of radically different flight behavior (e.g., a wartime surge).

These proportional models compute a historical cost per flying hour (CPFH) for a fleet of aircraft. To estimate future maintenance costs for that fleet, one multiplies the historical CPFH by the estimated number of flying hours. On a basic level, this idea makes sense: The more you operate any machine, the more likely it is to fail. Operating time is only one cause of failure, however. For aircraft, stresses from take-offs, landings, and other cycles also may cause failures. Failures may occur even when the aircraft is at rest; exposure to humidity, temperature, and dust can combine to make a part fail as the aircraft sits idle. These ground environment-related stresses may be more damaging than flight-related stresses. This fact has prompted some maintenance personnel to claim that aircraft can "heal themselves" by flying more frequently.

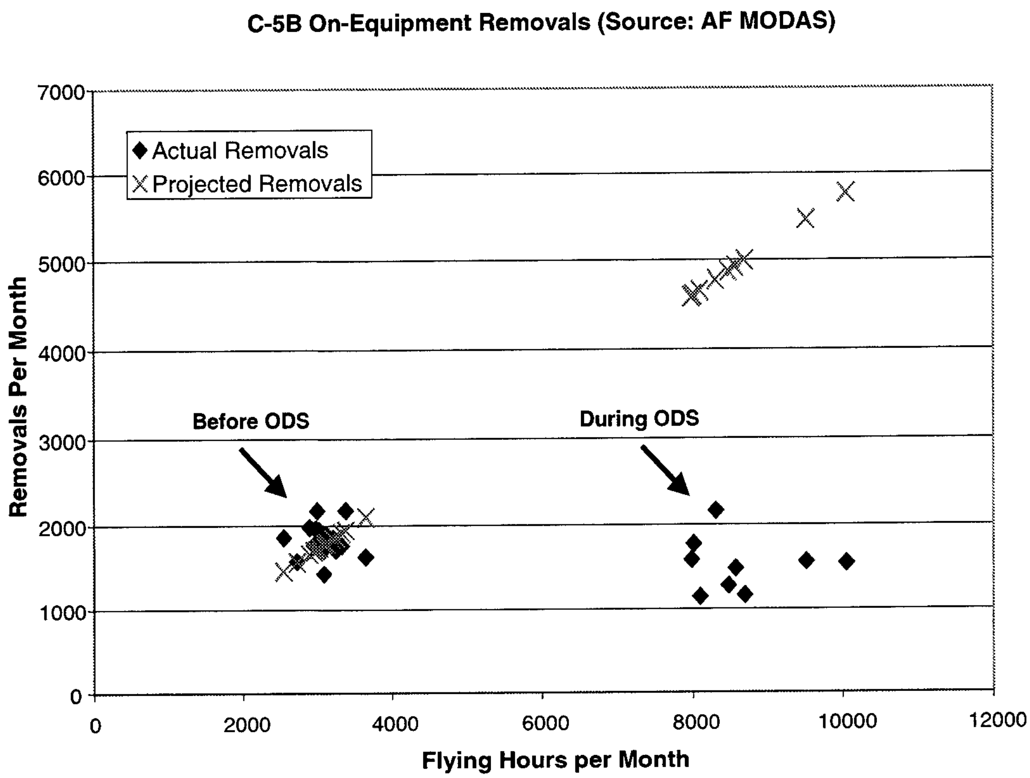
Why Does the Proportional Model Seem to Work?

If failures that are unrelated to flying hours outnumber those that are related to flying hours, why does the proportional model work at all? When nothing changes in the way an aircraft flies (and rests) from one period to the next, it is perfectly reasonable to use flying hours as a predictor for removal-causing failures. For that matter, it is equally reasonable to use possession hours or landings. This is because, when flight behavior does not change, the failure rate from each potential cause of failures remains constant. What happens, however, when flight behavior does change?

When the Proportional Model Fails

When a fleet of aircraft significantly changes its flight behavior from one time interval to the next, flying hours and the other factors that affect failures begin to diverge. In a typical surge, for example, flying hours increase dramatically, but landings remain about the same, and ground hours fall slightly. The proportional model would remain valid if flying hours maintain a strong correlation with removals during the surge. Unfortunately, they do not. How bad can the problem be? A proportional model predicted three times as many removal-causing failures for the C–5B during the Gulf War surge as actually occurred (see Figure 1-1).[1]

Figure 1-1. Projected and Actual C–5B Removals, Before and During Operation Desert Storm



For the C–5B, the Gulf War surge prompted a sharp reduction in the number of landings per sortie and a sharp increase in sortie duration. Because flying hours tripled, the number of events that can lead to an operational failure also tripled. The number of landings remained constant, however, and the number of ground hours was *smaller*: More flying hours mean less ground hours. Therefore, the surging C–5B fleet experienced fewer events that could cause cycle-induced or ground-induced failures. Because these events remained constant, overall removals during the surge remained roughly at pre-surge levels.

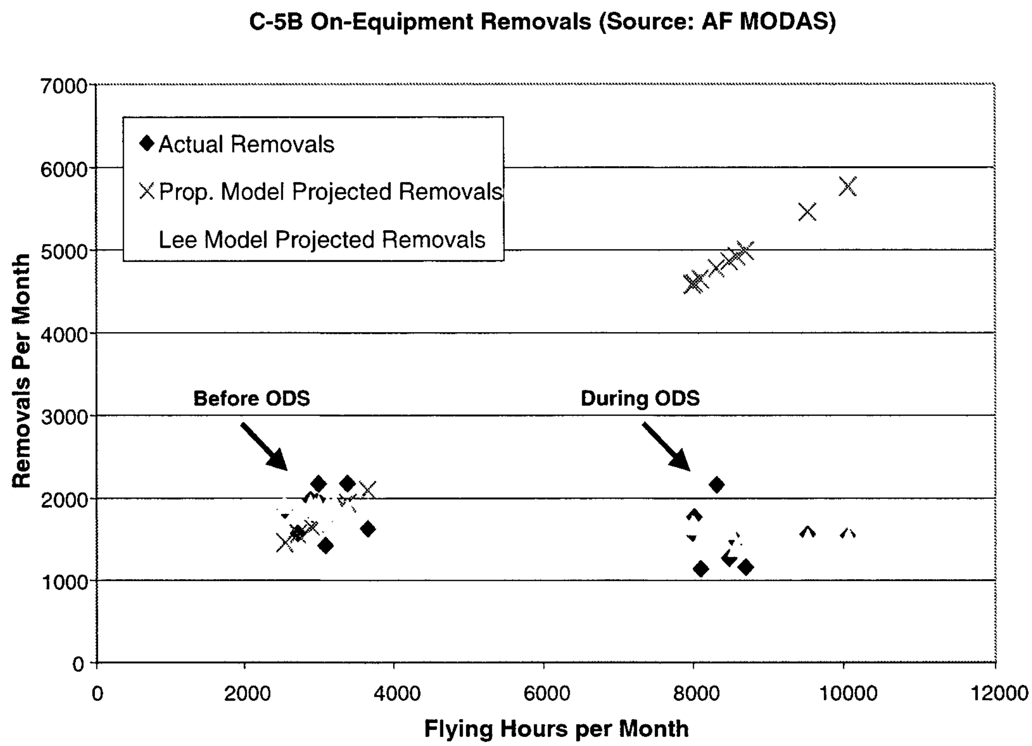
A PHYSICS-BASED ALTERNATIVE

To be consistently accurate, a material consumption model must consider more parameters than flying hours. It also must consider ground time and major stress-inducing cycles. David Lee created a physics-based model that considered the ground environment, flying hours, and take-off/landing cycles to predict

[1] We shorten "removal-causing failures" to "removals" for the remainder of this report.

removals.[2] He applied it to the same Gulf War C–5B data that confused the proportional model. Lee's model provided more accurate results than the proportional model for the Gulf War surge (Figure 1-2).

Figure 1-2. Actual Removals with Proportional Model and Physics-Based Model Developed by Lee



Lee's approach clearly shows promise. The question that motivated this study was this: Is implementation of this model throughout the Air Force feasible? The answer is yes, if the following premises apply:

- The model applies to many aircraft and scenarios, not only to the C–5B in ODS.
- Analysts can obtain data required to calibrate the model.
- Analysts can create or apply a simple computer tool to manage the data and calibrate the model.
- Analysts can easily review and interpret model results.

[2] David Lee, *The Cost Analyst's Companion*, (McLean,VA: LMI 1997).

In the remainder of this report, we describe the initial and final formulation of our physics-based model and outline the method we used to validate it. We then present one sample analysis each for a transport, fighter, and tanker. The examples show that our model performs better than the proportional model.

Chapter 2
Development of the Model

We used Lee's physics-based removal model as the foundation for our work.[1] Lee's model considers three input parameters:

- Take-off/landing cycles
- Ground hours
- Flying hours.

Other operating cycles may cause cycle-induced removals, but the model assumes that their contributions are small.

The probability P_c of a cycle-induced removal is

$$P_c = 1 - (1 - P_t)(1 - P_l), \qquad \text{[Eq. 2-1]}$$

where P_t is the probability of a failure during takeoff and P_l is the probability of a failure during landing. Cycle-induced failures follow a binomial distribution; that is, the probability of m_c failures in N cycles is $B(m_c, N, P_c)$.

Lee proposes that flight-induced removals and ground-induced removals both follow a Poisson process. Under this assumption, the number of flight-induced removals produced in time t_f is a discrete Poisson distribution with parameter $\lambda_f t_f$. Similarly, the number of ground-induced removals produced in time t_g has a discrete Poisson distribution with parameter $\lambda_g t_g$.

If the number of cycles, ground hours, and flying hours each is sufficiently large to produce more than 100 removals, we can approximate each of the three distributions above with a normal distribution. Using this assumption, the model predicts that the removals will have a mean of $NP_c + \lambda_f t_f + \lambda_g t_g$ and a variance of $NP_c(1 - P_c) + \lambda_f t_f + \lambda_g t_g$.

[1] David Lee, *The Cost Analyst's Companion,* (McLean,VA: LMI 1997).

THE CURRENT PHYSICS-BASED MODEL

Further investigation prompted the research team to make two basic modifications to the original model:

- We changed the ground-induced removals mechanism from a Poisson process to a binomial process. The use of a binomial process here presumes that ground stresses—mainly related to temperature and humidity—occur in a daily cycle. A Poisson process assumes that this environment places a more-or-less constant stress on the aircraft. To convert ground hours to daily ground cycles, we divided ground hours by 24.
- We separated cold cycles (the initial take-off and final landing of a sortie) from warm cycles (touch-and-goes). The distinction makes little or no difference in fighters because they seldom perform touch-and-gos, regardless of their mission. Transports, however, may perform several take-offs and landings per sortie during normal or training operations. Transports make significantly fewer warm cycles during a surge, however. Separating cold cycles from warm cycles allows for the possibility that the two events produce significantly different types and levels of stresses. More important, the separation shows the strong relationship between cold cycles and flying hours in tankers and transports.

For the aircraft we discuss in this report, the original and modified models produce nearly identical results, although the mix of removals differs. The original model tends to predict that the ground environment is the dominant cause of removals. In fact, it often predicts that the ground environment causes 75 percent or more of the total removals. The new model places ground-induced removals between 19 and 53 percent of the total. We believe that although the ground environment is significant, it is not as significant as the original model predicts.

CRITICAL PARAMETERS

With these modifications, the four critical factors affecting removals are

- ground hours ÷ 24 (ground *cycles* or ground *days*),
- flying hours,
- warm take-off/landing cycles (warm cycles), and
- cold take-off/landing cycles (cold cycles).

Ground Days

Let the probability that a ground cycle will cause a removal be P_g. Let N_g be the number of ground days when a removal can occur. ($N_g = t_g/24$, where t_g is the number of ground hours.) If the failure mechanism follows the normal approximation to the binomial distribution, then the mean number of ground-induced removals is N_gP_g. The variance is $N_gP_g(1 - P_g)$.

Flying Hours

This model treats flying hour-induced removals the same way Lee's original model does. The number of flight-induced removals produced in time t_f is a discrete Poisson distribution with parameter $\lambda_f t_f$. Using the normal approximation to the Poisson distribution, the mean and variance of flying-hour-induced removals both remain $\lambda_f t_f$.

Cold Cycles

With each sortie, the plane must start up its engines and other flight systems, take off, land, and shut down. The start-up and shutdown processes introduce stresses on the plane that may not be present in a warm (touch-and-go) take-off/landing cycle; therefore, we separate these cycles from other landings and take-offs that may occur during a sortie.

Each sortie begins with one start-up and take-off and ends with one landing and shutdown. The number of sorties is equivalent to the number of cold cycles. Because each cycle is a potential removal-causing event, we model cold cycle-induced removals with the normal approximation to a binomial distribution. If N_{cc} is the number of sorties and P_{cc} is the probability of a removal per cold cycle, the mean number of these removals is $N_{cc}P_{cc}$; the variance is $N_{cc}P_{cc}(1-P_{cc})$.

Warm Cycles

Warm cycles correspond to pairs of take-offs and landings that occur during a sortie. Although they may not be as stressful as cold cycles, they may be a significant cause of removals. This is especially true for transports; these planes may average three or more warm cycles per sortie during peacetime. Fighters, on the other hand, may have only a handful of warm cycles in a squadron for an entire year. In this case, these cycles should be excluded from the model.

We also model warm-cycle-induced removals with the normal approximation to a binomial distribution. If N_{wc} is the number of sorties and P_{wc} is the probability of a removal per warm cycle, the mean number of these removals is $N_{wc}P_{wc}$; the variance is $N_{wc}P_{wc}(1 - P_{wc})$.

Model Mean and Variance

The mean and variance of total removals results from combining the means and variances of each model parameter independently:

$$\mu = N_g P_g + N_{cc} P_{cc} + N_{wc} P_{wc} + \lambda_f t_f . \qquad \text{[Eq. 2-2]}$$

$$\sigma^2 = N_g P_g (1 - P_g) + N_{cc} P_{cc} (1 - P_{cc}) + N_{wc} P_{wc} (1 - P_{wc}) + \lambda_f t_f . \qquad \text{[Eq. 2-3]}$$

A careful look at these equations shows one important constraint: because $N_{wc} = N_{landings} - N_{sorties}$ and $N_{cc} = N_{sorties}$, it is possible to express the model mean as

$$\mu = N_g P_g + N_{sorties} \left(P_{cc} - P_{wc} \right) + N_{landings} P_{wc} + \lambda_f t_f . \qquad \text{[Eq. 2-4]}$$

A negative probability of failure caused by sorties does not make sense. Thus, P_{cc} must be constrained so that it is always greater than or equal to P_{wc}.

Implementing the Model

To calibrate the model, the analyst uses a set of calibration data that includes a set of time intervals and the removals, sorties, ground hours, flying hours, and landings that occur in each. The model computes a predicted set of removals for each of these intervals, and an optimization routine computes the best fit between predicted and actual removals. The model variances vary from interval to interval (i.e., they are *intrinsically heteroscedastic*), so we compute a maximum likelihood estimator numerically, rather than by a least mean squares fit. (Lee provides a mathematically rigorous discussion of the method; see Appendix.) The optimization scheme may seem complex, but it is simple to set up and run on a single Excel spreadsheet, using the optimizer add-on.

Proper Use of the Model

As with most physics-based models, the analyst must use it with appropriate care. First, the calibration set should consider how to treat highly collinear parameters. If two parameters are indeed collinear, the analyst should consider simplifying the model to eliminate one of them or use a more sophisticated method, such as ridge regression, to determine the principal components of the model. A mathematically rigorous method to diagnose collinear parameters is beyond the scope of this project. (It is sufficient to make this determination through visual inspection.)

Second, the analyst should select time intervals that are large enough to ensure that each parameter will produce at least 100 removals. Because the model is a linear combination of the four parameters, sorting predicted removals by their underlying cause is easy, provided all four parameters are independent.

Third, the analyst must consider slow-developing, time-dependent trends in the data. For example, three of the four aircraft we discuss in this report exhibited a *removals bathtub*. In the first few years of service, a new fleet often shows a gradual decrease in removals that cannot be explained by changes in flight characteristics. As the fleet matures, this trend levels out. Beyond a certain age, removals increase again. This trend can last for several years. Neither the physics-based model nor the proportional model considers removals bathtubs. Failure to remove these bathtubs from the data may reduce the accuracy of both models.[2]

Ensuring Independent Parameters

If a fleet's flying behavior varies little over time, the sortie duration, number of landings per sortie, and number of flying hours per month will remain essentially constant. In this case, discerning the independent effect of each parameter is impossible, and using one variable to predict removals is equivalent to using all of them. When flying behavior changes, the parameters may respond independently. Discerning the individual effects of each is then possible.

Even when flying behavior changes significantly, at least two of the four parameters may still behave similarly. In this case, we gain nothing by keeping both parameters in the model. Because the two parameters are redundant, treating them as separate inputs can be misleading: Although an optimizer will arbitrarily assign failure probabilities to each, only their combined effect has meaning. In this situation, we chose to eliminate one of the redundant parameters. This approach makes the results easier to interpret. It also keeps the model as simple as possible, thereby minimizing the computational load. Finally, by preserving the independence of the model's parameters, we can enumerate predicted removals by their source. This strategy enables us to see whether the normal approximation is valid for the selected calibration set.

For the purposes of this study, a visual inspection is sufficient to check the parameters for collinearity. To make this inspection easier, we scale each of the four parameters before plotting them. For example, if a fleet of aircraft flies an average of 2,000 hours for a set of data and sees 1,000 ground cycles, ground cycles must be multiplied by 2 to scale them to flying hours.

Our research has shown that cold cycles and flying hours often exhibit collinearity when they are plotted this way (see Figures 2-1 and 2-2). The C–17 fleet's flying hours and cold cycles clearly are related; the F–16Cs at Aviano Air Base exhibit a departure between these two parameters.

[2] We discuss removals bathtubs in greater detail in Chapter 3.

Figure 2-1. Strongly Related Flying Hours and Cold Cycles

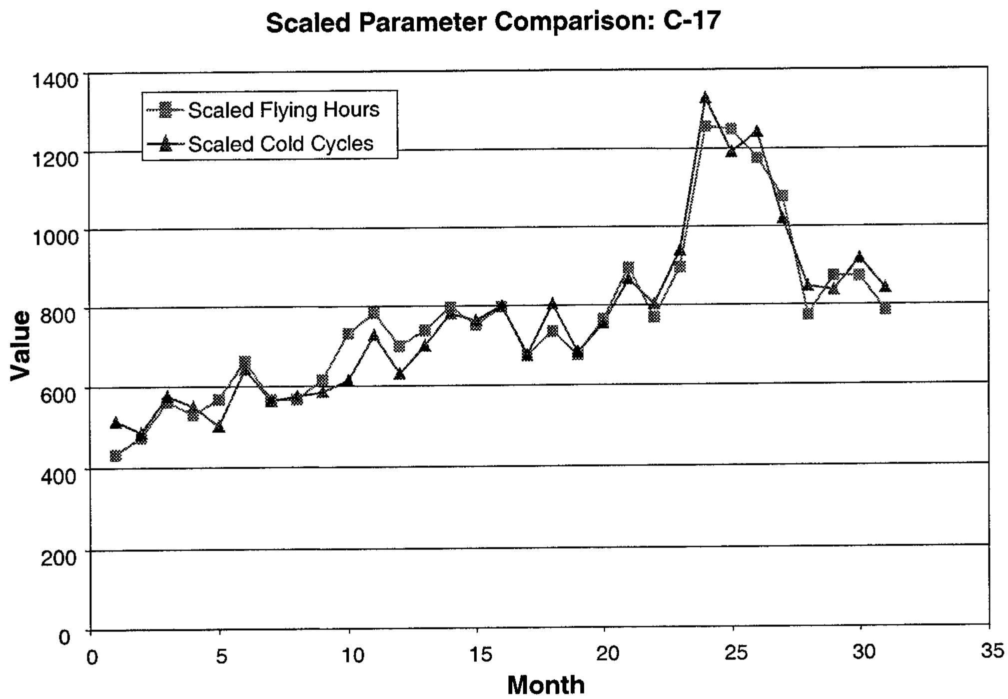
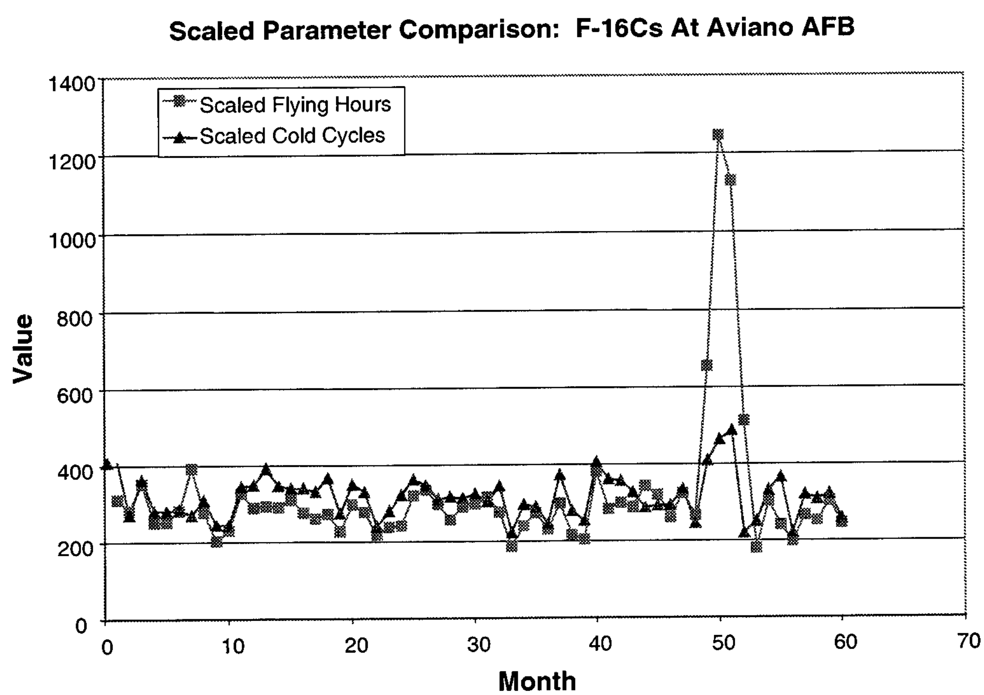


Figure 2-2. Independent Flying Hours and Cold Cycles



The tankers and transport planes we studied show a strong relationship between flying hours and cold cycles; therefore, we chose to eliminate flying hours from the model when we analyzed the C–5B, C–17, and KC–10. Eliminating cold cycles would have been equally valid, but we retained cold cycles over flying hours for two reasons:

- We believe take-offs and landings are more jarring than steady-state flying.
- We want to show that a model can predict removals accurately without using flying hours.

The F–16C showed very few warm cycles. Although these cycles are not collinear with other variables, we eliminated them as a parameter because they were too few to be a significant source of removals. Inspection of other fighter aircraft data reveals that they make significantly fewer touch-and-goes than transports, which leads to the conclusion that cold cycles should be eliminated in modeling all fighters.

The limited scope of this study prevents us from taking a more than cursory approach to the issue of customizing the model for specific types of aircraft. With further study, we would consider this issue with more scientific rigor. For the purposes of this study, however, eliminating flying hours from the transports and tankers and eliminating warm cycles from the F–16Cs makes the model easier to calibrate and evaluate.

Choosing a Proper Time Interval

The analyst must consider one major trade-off in selecting the time interval. Too small a cycle may provide too few removals to make the normal approximation accurate. For example, if monthly data produce fewer than 100 ground cycle-induced removals, the analyst should combine the data into bimonthly or even quarterly time intervals. Too large a cycle may cancel important variations in the data through an averaging effect. For example, suppose a group of aircraft undergoes an intense, 3-month surge followed by a 3-month period of reduced flying hours. With a 6-month time interval, these significant changes in flight program might become indistinguishable from 6 months of normal flying.

If the time intervals cannot be set so that each parameter accounts for 100 removals per interval, the analyst should consider the following alternatives:

- Use the binomial or Poisson distribution, as appropriate, rather than the normal approximation, for any parameter that fails to meet the criterion.
- Eliminate the parameter if its contribution to the total removals is insignificant.
- Use the normal approximation and accept a slightly less accurate result.

The fleets that we investigated all produced a sufficient number of removals to allow the use of the normal approximation with monthly time intervals.

Chapter 3

Other Modeling Issues

The physics-based model accurately predicts removals, given physical data about the number and types of take-offs and landings, the number of 24-hour ground cycles, and flying hours. The model assumes that the likelihood of failures from any one of these factors remains the same over time. The data show, however, that for most aircraft the likelihood of failure does not remain constant over long time intervals. Long-term trends—called *removals bathtubs*—vary the general likelihood of failure. These trends affect the physics-based model and the proportional model alike. In this chapter, we describe the removals bathtub and explain how to correct both models for it.

The first step in computing an accurate material consumption cost is to create a model that accurately predicts the number of removals. The second step is to identify any physically explainable trends in the data that the model doesn't treat, and then correct the model for those trends. For this reason, we correct for removals bathtubs. The third and final step is to convert the removals prediction into an aggregate cost for material consumption. Because of limitations with the data, we do not complete this step.[1] In this chapter, we provide a general framework for calculating a cost from predicted removals.

SLOW-DEVELOPING TRENDS—REMOVALS BATHTUBS

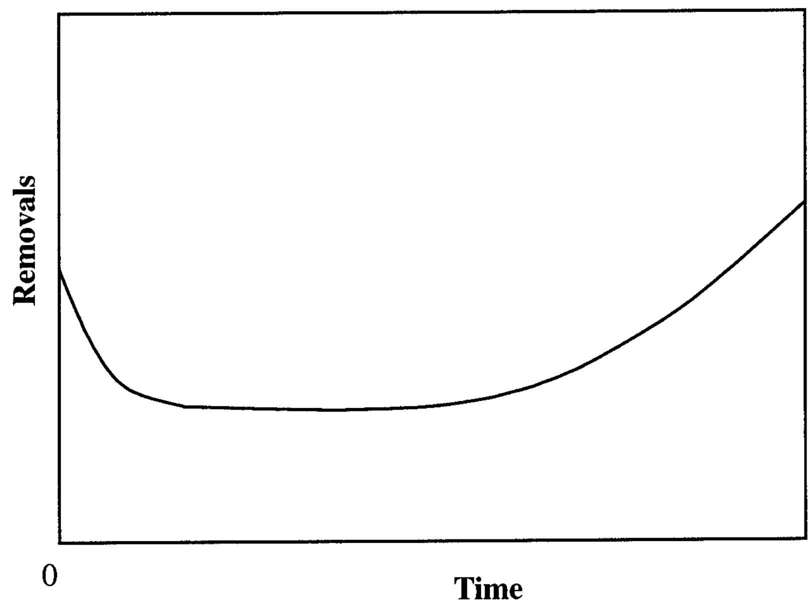
Figure 3-1 demonstrates the concept of a removals bathtub. For the first few years, some fleets go through a period of declining removals. Maintenance chiefs for the aircraft we discuss in this report suggest several reasons for this decline:

- Maintenance crews gain experience
- Manufacturers gain experience (when delivery is spread over several production lots)
- The maintenance organization relaxes requirements after a "burn-in" period
- Better written technical manuals result in less wasteful maintenance.

As the aircraft ages, removals creep up again—possibly because clearances and voltages drift away from optimum values. Clearances and other such parameters change over time as a result of cumulative wear, fatigue, and corrosion. These problems force components to work harder, which hastens their aging.

[1] Data required for this calculation will be available soon.

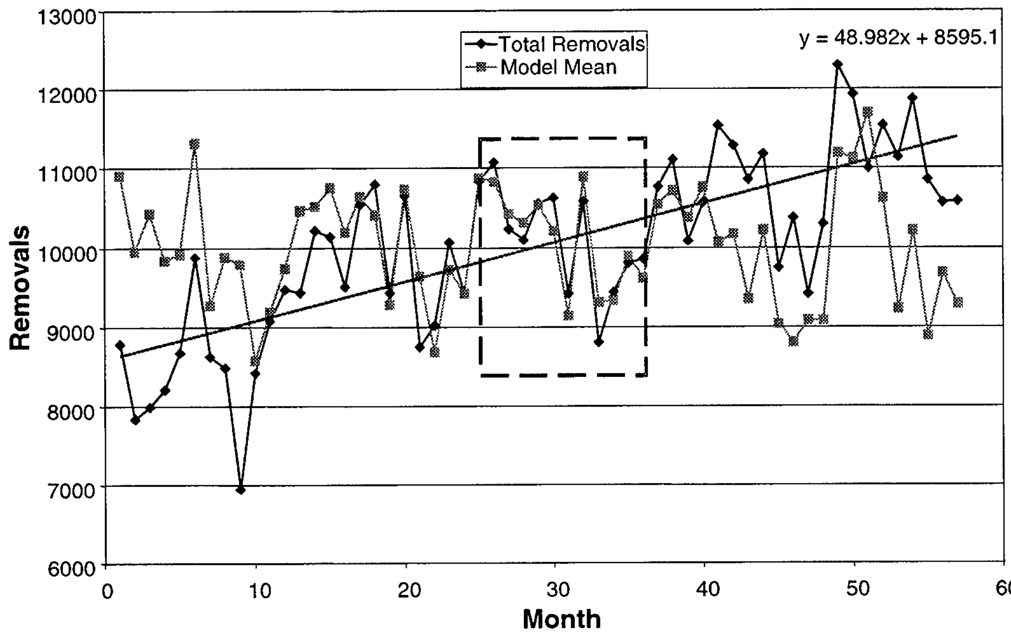
Figure 3-1. Removals Bathtub (exaggerated)



Although the combination of causes may vary from fleet to fleet, we often see a gradual rising or falling trend in overall removals over several years. Neither our physics-based model nor a proportional model can explain these trends; both assume that the aircraft's reliability remains constant throughout the calibration period and the estimation period. If the estimation period occurs several years after the calibration period, this assumption may not be valid.

For example, the F–16C fleet exhibits evidence of an upward trend in removals that cannot be explained by an increase in possession time, flying hours, or takeoffs and landings (Figure 3-2). As a result, a physics-based model produces a relatively good fit in the calibration region (months 25 to 36, as indicated by the box), but its accuracy becomes progressively worse in either direction from the calibration set. Although the model captures the transient variations in the data, it cannot predict the underlying steady-state increase in removals. The F–16 maintenance chief confirmed that the aging fleet has had rising removal rates for the past several years because of its increasing age.

Figure 3-2. Data Fit with Gradually Rising Trend in Removals (calibration period: months 25–36)



After such a trend is confirmed, the best response is to remove the trend from the data, fit the model to the modified data, and apply the trend to the resultant model.[2] If fatigue is the cause for the rise in removals, the data should follow a slow exponential trend. (The true nature of this trend should be confirmed through further study.) A gradual exponential can be approximated reasonably by a straight line. This approach greatly simplifies the math required to transform the model to fit the underlying trend. When a linear approximation is appropriate, the following steps will transform the model:

1. Fit a trend line to the actual removals data and note its slope.

2. Perform a *shear transformation* on the actual data to remove the trend.

$$\begin{aligned} Mon' &= Mon \\ R' &= R - (Mon - Mon_c)m \end{aligned} \qquad \text{[Eq. 3-1]}$$

Mon is the number of the current month (or the number of the time intervals chosen). This number remains constant under the transformation. Mon_c is the midpoint at the center of the calibration set. In this case, $Mon_c = (25 + 36)/2 = 30.5$. R is the number of removals for that month. The last term, m, is the slope of the trend line.

3. Calibrate the model using the transformed actual data.

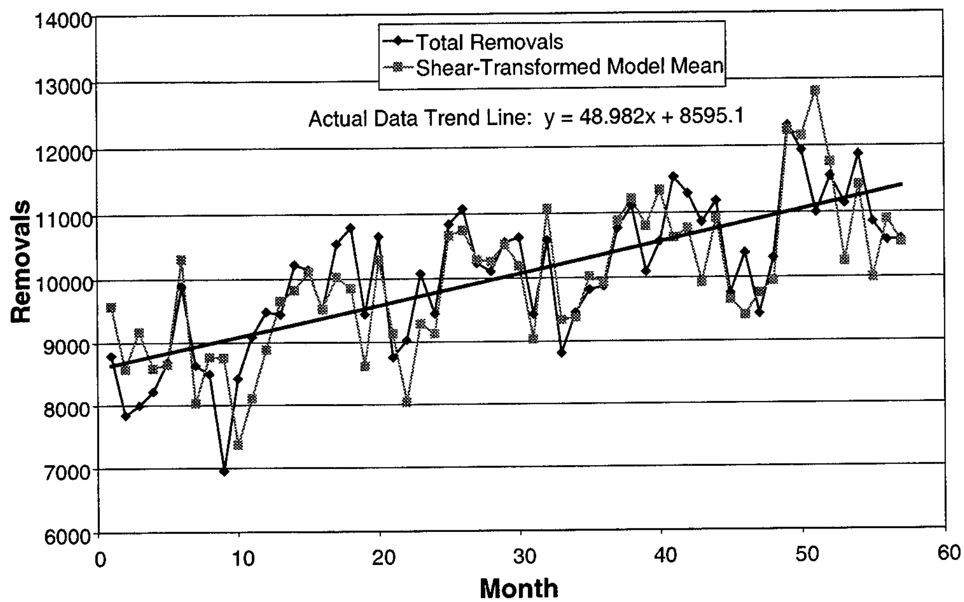
[2] George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, *Time Series Analysis: Forecasting and Control*, 3rd Edition (New York: Prentice Hall, 1994).

4. To compute predicted removals from a set of known model parameters, compute the model mean, then transform it using the following set of equations:

$$\begin{aligned} Mon' &= Mon \\ R' &= R + (Mon - Mon_c)m \end{aligned} \qquad \text{[Eq. 3-2]}$$

Figure 3-3 shows how the model fits the F–16C data with this process.

Figure 3-3. F–16C Data Fit, Accounting for Underlying Rising Trend



This technique will improve the accuracy of this model and a proportional model alike. An analyst should be careful to identify and account for these trends, regardless of the model the analyst uses.

ESTIMATING INPUT PARAMETERS

To use this model, the analyst must be able to provide reasonable estimates for each of the input parameters that feed it. Analysts project flying hours now; with the new model, they would need to continue to do so. Ground cycles are directly related to ground hours at an operational unit. This value is as easy to estimate as flying hours: If one knows a fleet's expected flying hours and the number of aircraft on hand, one also knows the fleet's expected ground hours. Predicting landing cycles should be relatively simple too, if the analyst knows how many sorties the aircraft will fly in the given duration, what types of missions are expected, and how many landings occur in a typical sortie of that mission type. One can extrapolate these data from the Reliability and Maintainability Information System (REMIS) database. An analyst with knowledge of the operational characteristics of a given aircraft may be able to make a reasonable estimate solely on the basis of experience.

RELATING REMOVALS TO COST

With a proportional model, computing a cost is trivial. By definition, the CPFH model directly relates cost and flying hours. With a physics-based model, relating the physical events that the model predicts (removal-causing failures) and their cost is not as easy. For example, if a specific removable item costs 10 times as much to replace as another and the model considers all removals to be the same, how do we compute a cost?

The simplest way to do this calculation is to compute an average cost for a removal and then multiply all of the predicted removals by that value. A CPFH model implies this technique:

$$(FH)\times\left(\frac{removals}{FH}\right)\times\left(\frac{\text{cost}}{removal}\right)=(FH)\times\left(\frac{\text{cost}}{FH}\right)=\text{cost}\,. \qquad \text{[Eq. 3-3]}$$

A better way to compute a cost would be to calibrate the model to predict the number of specific assemblies, subassemblies, or parts removed, rather than the number of all removals. The model can predict the removals of a specific piece of equipment just as easily as it can predict general removals; the analyst must simply substitute equipment-specific removals data for general removals data when calibrating the model. The REMIS database provides this equipment-specific removals data. Therefore, the analyst can

- predict the number of each part that will be removed, using projected flying hours, possession hours, and landing data for a given aircraft;
- multiply the number of each part's removals by the average cost to replace that part; and
- add the individual costs together to compute a total cost of spares for a coming period.

More realistically, an analyst will run the model for the 10 or 20 items that are most costly to remove. After performing the preceding procedure to compute a combined cost for those parts, the analyst would then scale that number by a certain percentage to compute a total spares cost. For example, if 90 percent of the cost of spares for a particular aircraft comes from 15 parts, the analyst would obtain individual removals data on those 15 parts. Then the analyst would create a model for each of those parts, use those models to compute a projected cost for each, sum them, and multiply by 10/9 to compute a total projected cost for spares.

This modeling method assumes that the removed item is also replaced. Because items often are removed for repair and then reinstalled, the analyst must distinguish between items removed for replacement and items removed for repair. Spares consumption data should relate more directly to costs than removals. The model can be adjusted easily to predict consumption cost rather than removals. Although we were unable to obtain this type of data for this study, we anticipate that new Air Force Total Ownership Cost (AFTOC) data will have this information.

Chapter 4

Methods and Results

For the model to be generally useful, it must apply to several types of military aircraft. For brevity, we discuss the results we obtained in fitting the model to a representative transport plane (C–17), fighter (F–16C), and tanker (KC–10). We selected each of these planes to represent their class because they experienced at least one sustained surge during recent years. These surges most likely were caused by the Kosovo air campaigns. These surges are much smaller than that shown in the ODS-era C–5B data, so the improvement over the flying-hour model will be correspondingly less dramatic. The physics-based model still noticeably outperforms the proportional model in each case.

Before we examine the Kosovo-era data, we examine the ODS C–5B data that motivated this study. The C–17, KC–10, and F–16C data follow. We analyzed each data set with the same approach:

- We scaled each individual parameter to identify any strong relationships between them. We visually identified collinear parameters by using a graph of the scaled parameters. We considered a more mathematically rigorous approach, such as correlation coefficients, but decided that visual identification was sufficient given that this research is a feasibility study, not an advanced investigation.
- When a strong relationship existed between two or more parameters, we chose to eliminate redundant parameters. The analyst can choose a method such as ridge regression that will allow all of the parameters to remain in the model. Given the scope of this study, however, we decided against using such advanced techniques.
- We compared the physics-based model and the proportional model graphically, using the remainder of the data set, and comment on the results.

We used four different calibration sets for each aircraft in the study. For the first three calibration sets, we divided the data roughly into thirds to see how well the physics-based model and the proportional model performed. We also chose a fourth calibration set, which also included approximately a third of the data for each aircraft. This fourth set included the months before the surge, along with roughly the first half of the surge data. This set is the "best" calibration set for both models, given the constraint of leaving roughly two-thirds of the data as test data. It is best for the physics-based model because it guarantees that the aircraft undergo at least two distinct operational patterns. If the aircraft never change their operational pattern, all of the parameters will be roughly proportional to time, and discerning their individual affect on removals is difficult. The calibration set is best for the proportional model because it includes high-flying hour, low-removals surge data with data from typical peacetime operations. As a result, the proportionality constant is smaller, which means that the model will not overpredict the surge time removals as grossly as it might if surge data were not included in the calibration. Moreover, because the set includes non-surge data, the proportionality constant does not become so low that the model will grossly underpredict removals that occur during peacetime operations.

We call this fourth data set a "leading-edge" data set because it includes the first half of the surge and the months that lead up to it. In the discussion that follows, we focus on the leading-edge data set for two reasons: This set best represents the spectrum of flying behaviors for the aircraft and the proportion of time spent doing each. It also provides the toughest test for the physics-based model because the proportional model tends to be most accurate for this calibration set.

In addition to showing the leading-edge calibration sets for all four aircraft, we show the C-5B models calibrated on the first third of the data. This set most clearly shows the disparity between the physics-based and proportional models that motivated this study. We also show all four calibrations of the F-16Cs stationed at Aviano Air Base to provide an example of how the models vary with the choice of calibration data. For the sake of brevity, we do not show the rest of the calibrations. We show all four sets of the Aviano F-16Cs because this aircraft displays the strongest relationship between flying hours and removals; therefore, if the proportional model can outperform the physics-based model, it will do so on one of these four data sets.

We used two statistics to describe the differences between each model and actual removals: the mean relative error and the root mean squared (RMS) relative error during surge months. The relative error of a data point i is

$$E_{rel\,i} = \frac{P_i - O_i}{O_i}, \qquad \text{[Eq. 4-1]}$$

where P_i is the predicted value for point i and O_i is the observed value at point i. The mean relative error of n data points is

$$M_E = \frac{\sum_{1}^{n} E_{rel_i}}{n}. \qquad \text{[Eq. 4-2]}$$

Computing the RMS error of n data points involves squaring the relative error of each point, taking the mean of the squared errors, and taking the square root of the mean:

$$RMS_E = \left(\frac{\sum_{1}^{n} E_{rel_i}^2}{n} \right)^{1/2} . \quad \text{[Eq. 4-3]}$$

Relative error measures how well the model matches the central tendency of the observed data. It preserves the sign of the difference between a predicted and actual value; therefore, if the model overpredicts some data and underpredicts others, the errors tend to cancel. A good model should match the central tendency of the observed data closely, but it also should predict local peaks and valleys in the data. RMS error is a measure of the average magnitude of the difference between each predicted and actual value. It shows how well the model predicts these local data variations. If the model consistently over- or underpredicts the data in a set, the relative error and RMS error will have approximately the same magnitude.

Table 4-1 summarizes the results of each of the 16 model calibrations for the proportional model and the new physics-based model. We discuss the initial physics-based model (the Lee model) in the C–5B section for historical purposes, but we do not include it in the table or in the discussions of the other aircraft fleets.

OPERATION DESERT STORM: C–5B

The research team obtained ODS-era data from the Air Force's Maintenance and Operational Data Access System for the months of March 1989 to March 1992, inclusive. These 37 months of data show a surge for the months of August 1990 to May 1991 (see Figure 4-1). During this time, flying hours typically were double or triple the mean value before the surge. As a result, the proportional CPFH model fails dramatically. Can the new model do better? To answer this question, we evaluate each model's performance individually, restricting our attention to the mean relative errors for several regions of the data.

Figure 4-2—a graph of all four model parameters scaled to flying hours—shows a strong topological similarity between flying hours and cold cycles. Although the two variables have slightly different means in the pre-surge, surge, and post-surge regions of the graph, they clearly are closely correlated. As we discuss in Chapter 3, we eliminated flying hours from the model for this reason.

Figure 4-1. C–5B Flying Hours Before, During, and After Operation Desert Storm

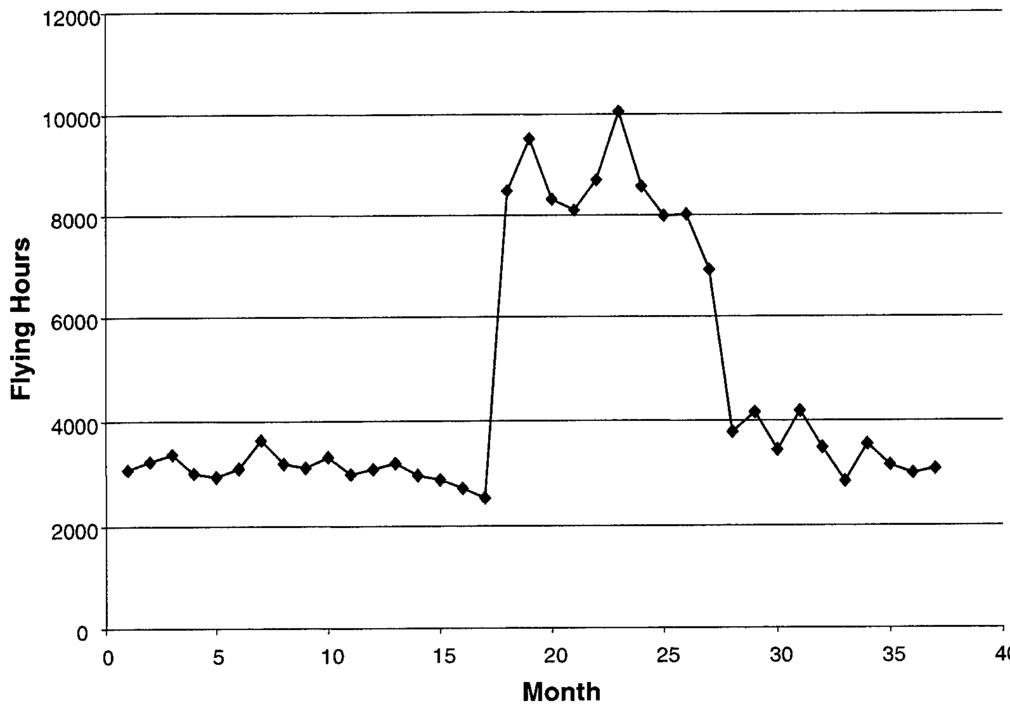
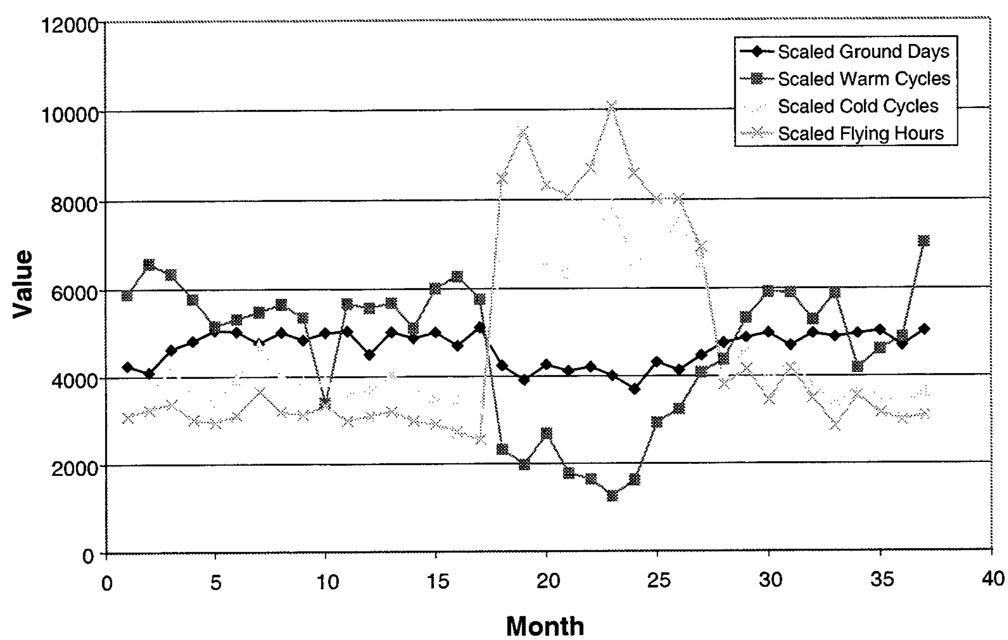
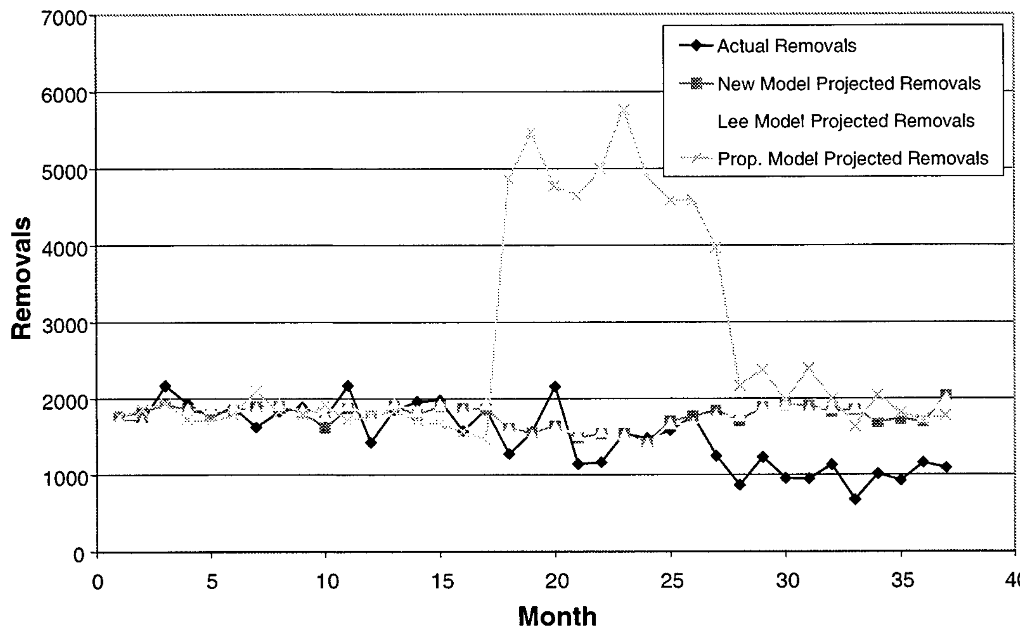


Figure 4-2. All Parameters Scaled to Flying Hours, Showing Strong Relationship Between Cold Cycles and Flying Hours



Cold cycles, warm cycles, and ground cycles appear to be uncoupled throughout the data set. Thus, we can calibrate the model anywhere in the data set and expect to see good results from a physics-based model. We describe two of the four calibrations we performed on this data: the calibration that used the first third of the data and the leading-edge calibration. The leading-edge calibration includes the 10th through the 21st months—the 8 months before the start of the surge and the first 4 months of the surge. For the C-5B only, we consider Lee's original physics-based model and the new model that refines it. (Comparison of the two shows why we chose the latter over the former.) For the other three aircraft discussed in this report, we show only the new physics-based model. Figure 4-3 shows the three different removal models overlaid on actual removals.

Figure 4-3. Removals Predictions for C–5B During Operation Desert Storm (calibration period: months 1–12)



Proportional Model

The proportional model follows the general trend in actual removals relatively well before the ODS surge. In the first 5 months following the calibration year, it underpredicts removals slightly, but not so much so that it differs significantly from the other two models. During the surge, however, the proportional model diverges from actual removals so badly that it dwarfs any other divergences in the graph. The mean relative error shows that the proportional model overpredicts removals for months 18–27 by 236 percent.

The proportional model also overpredicts removals by more than 100 percent for the post-surge months—although it does not significantly underperform the other two models for this period. Why do all three models perform so badly here? Several possibilities exist:

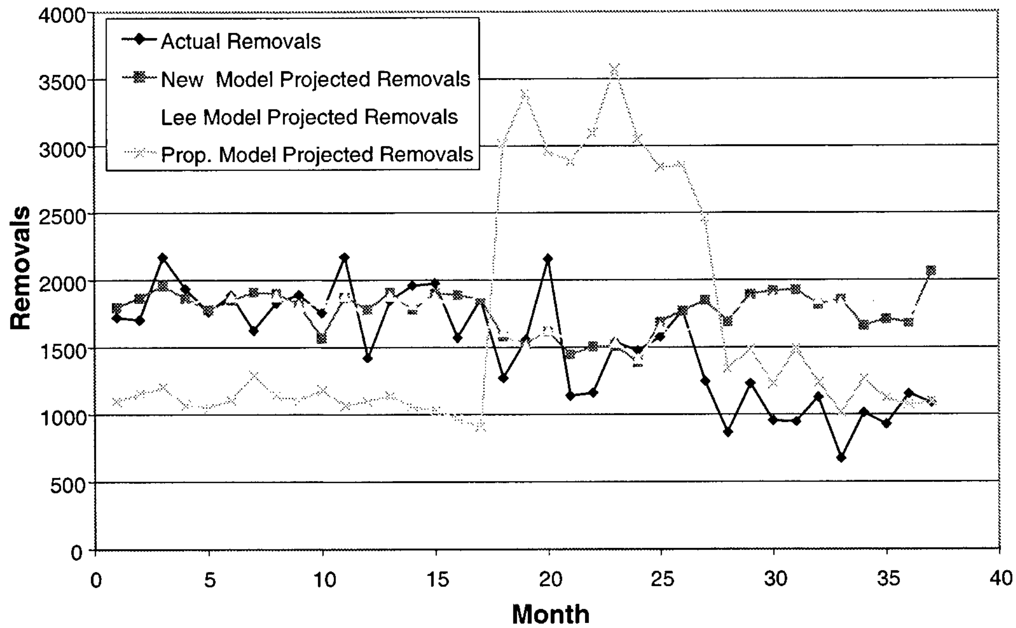
- The relatively young C–5B fleet may have been experiencing a "left side of the bathtub" reduction in removals, with the start masked by the wartime surge.
- The Air Force may have flown or serviced these aircraft in fundamentally different ways before and after their participation in the Gulf War. This difference may be related to a physical effect that is not captured by the parameters of any of the models.
- Removals may have been underreported, possibly because of a change in reporting policy or a shift to a new maintenance reporting system.

Because of the age of the data, we cannot ascertain which of these causes contributed to the unusually low post-ODS removals. The C–5B maintenance personnel we interviewed could not positively identify any one of the foregoing possibilities as the true cause. Underreporting seems unlikely. A prevailing theory suggests that no significant spike in removals occurred during the Gulf War because a large amount of maintenance was deferred. If that were true, removals levels would be elevated after the surge even with significant underreporting. In reality, removals are depressed. It also seems unlikely that the planes flew radically differently after the surge than they did before and during the surge. Neither the data nor the maintenance personnel we interviewed suggest this possibility. The remaining possibility—a removals bathtub—seems to be the most likely scenario. Although the surge obscures the trend, there is a gradual decline in removals across the entire data set. We investigate this possibility later in this section.

The mysterious post-surge decline in removals aside, the proportional model fails spectacularly when we use the first third of the data as the calibration set. If we include some of the surge in the calibration, we could expect to do better. The leading-edge calibration set tests that hypothesis.

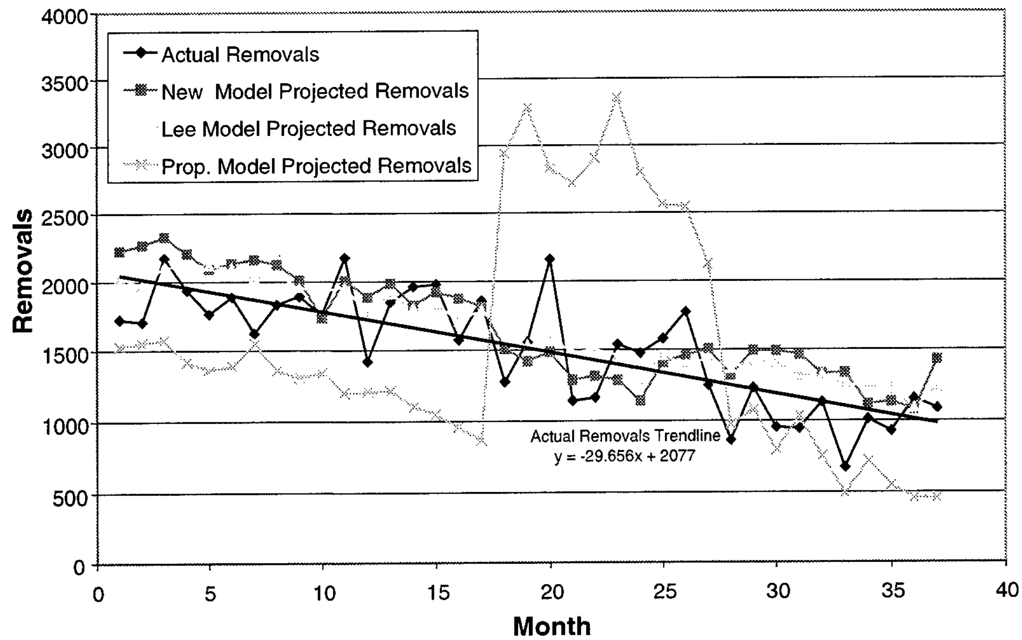
The proportional model fails for the leading-edge calibration as well (see Figure 4-4). The mean relative error shows that the proportional model overpredicts removals for the 10 months of the surge outside the calibration data by 109 percent. The disagreement between the model and the data is much smaller than for the previous calibration because it includes a few months of surge data in the calibration set, which effectively shifts the entire model down. Unfortunately, this same effect causes the model to underpredict the months before the surge: The mean relative error is –39 percent. The model does manage to fit the post-ODS removals better than the physics-based models; it overpredicts post-surge removals by 26 percent. This result is just a happy accident. The proportional model grossly underpredicts the data before the surge, and one expects it to grossly underpredict the data after the surge as well. Because the post-surge removals are much lower than expected, the model manages to fit the data reasonably well.

Figure 4-4. Removals Predictions for C–5B During Operation Desert Storm (calibration period: months 10–21)



The proportional model does work better when we remove the bathtub trend from the data (see Figure 4-5). Before the surge, the proportional model underpredicts the actual data by an average of 29 percent per period. During the surge, it over-predicts the data by 94 percent. After the surge, it underpredicts the data by 25 percent. Treating the removals bathtub improves the model before and during the surge by 10–15 percent. The model does not improve after the surge—but only because the bathtub in the untreated case happened to coincide with an expected underprediction. Overall, the fit is significantly better.

Figure 4-5. Fit of Models to C-5B Data After Treating Bathtub Trend (calibration period: months 10–21)



A return to the first calibration set also shows some improvement. Before the surge, the error slipped from –2 percent to –6 percent, but during the surge it improved from +236 percent to +203 percent. After the surge, the error improved dramatically: from +103 percent to +24 percent. The extremely high errors during the surge show that although we can expect better accuracy when we treat a removals bathtub, this improved method alone cannot compensate for the shortcomings of the proportional model. Physics-based models can correct these shortcomings.

Initial Physics-Based Model

To provide continuity to the study and show how it has evolved, we included the original removals model developed by Lee in the C–5B analysis. Like the current model, Lee's original model shows a major improvement over the proportional model. Lee's model does not vary significantly from the current one in overall trend; it does differ, however, in its underlying physics:

- Lee models ground environment-induced removals as a Poisson process, based on ground hours, rather than as a binomial process based on ground days.
- Lee's model includes flying hours. We have removed flying hours from the current model for the C–5B because of the strong correlation between flying hours and warm cycles.

- Lee lumps all take-off and landing cycles together; his model makes no distinction between cold cycles and warm cycles.

Lee's model has the following form:

$$\mu = \lambda_g t_g + N_c P_c + \lambda_f t_f \qquad \text{[Eq. 4-4]}$$

$$\sigma^2 = \lambda_g t_g + N_c P_c (1 - P_c) + \lambda_f t_f \qquad \text{[Eq. 4-5]}$$

λ_g and λ_f are the Poisson coefficients for ground-induced and flying-hour-induced removals, respectively. P_c is the probability of a removal because of a take-off or landing cycle.

Calibrating this model on the first year of data yields $\lambda_g = 0.0523$, $P_c = 0.0527$, and $\lambda_f = 0$. Lee's model predicts that flying hours have no effect on removals. In part, this occurs because landings and flying hours are somewhat correlated. Nevertheless, the results clearly demonstrate that ground hours and landings are major removal-causing factors: Ground hours account for 92.1 percent of the removals predicted, and landings account for 7.9 percent.

The physics-based model has a mean relative error before the surge of 2.1 percent—only slightly better than the proportional model. During the surge, however, the physics-based model remains accurate: It underpredicts the data by 22 percent, whereas the proportional model overpredicts the data by 203 percent. After the surge, the physics-based model also does slightly better: It overpredicts the data by 14 percent, whereas the proportional model overpredicts the data by 24 percent.

The second calibration set yields a slightly different set of results: $\lambda_g = 0.0455$, $P_c = 0.0766$, and $\lambda_f = 0.0303$. Thus, ground hours account for 80.2 percent of predicted removals, landings account for 11.7 percent, and flying hours account for 8.1 percent. The change in calibration region has only a minor effect on the outcome of the model, despite the major change in flying hours in the second calibration data set. This finding suggests that the physics-based model is more robust than the proportional model.

The physics-based model again significantly outperforms the proportional model. Pre-surge, its mean relative error is +7 percent, compared with –29 percent for the proportional model. During the surge, its mean relative error is +3 percent, compared with 94 percent; after the surge, its mean relative error is +32 percent, compared with –25 percent.

The New Physics-Based Model

Through comparison, we see that differences in assumptions between the current model and the original model have only a minor effect on the C–5B results. Both physics-based models are superior in accuracy to the proportional model, and both are more robust. The two physics-based models allocate predicted removals differently, however. Calibrated on the first year of data, the new model predicts that the ground environment accounts for 45.3 percent of removals, warm cycles account for 31.8 percent, and cold cycles account for 22.9 percent. These results differ significantly from those of the original model, although there is only a small difference in the predicted total outcome (Figure 4-5). The same is true for the second calibration set. The new model predicts that the ground environment accounts for 37.8 percent of the removals, warm cycles account for 37.1 percent, and cold cycles account for 25.1 percent.

The new physics-based model also is more accurate than the proportional model. For the first calibration set, the new model's mean relative errors are –3 percent pre-surge, –15 percent during the surge, and +36 percent post-surge. For the leading-edge calibration set, its mean relative errors are +12 percent pre-surge, –4 percent during the surge, and +9 percent post-surge. The original physics-based model predicts that the ground environment causes more than 80 percent of the removals in both calibration regions, yet the two models differ little in accuracy (Figure 4-5). Although we cannot discern which model more closely represents the physics involved (that is beyond the scope of this study), we feel that the original model places too much emphasis on the ground environment.

Neither physics-based model performs as well post-surge as pre-surge; the errors remain reasonably small, however. An additional issue, such as a major reduction in cargo weight, may be responsible for the additional reduction in removals. The C–5B has chronic reliability problems; therefore, one might reasonably speculate that they flew a more conservative flying regime after ODS to preserve their mission readiness. Although we could not confirm this hypothesis directly from C–5B maintenance personnel, F–16C and C–17 personnel suggested that cargo loading (or weapons configuration for fighters) can significantly affect removals. Although the C–5B data can do no more than raise these questions, we think investigating them further with more recent, complete data would be useful.

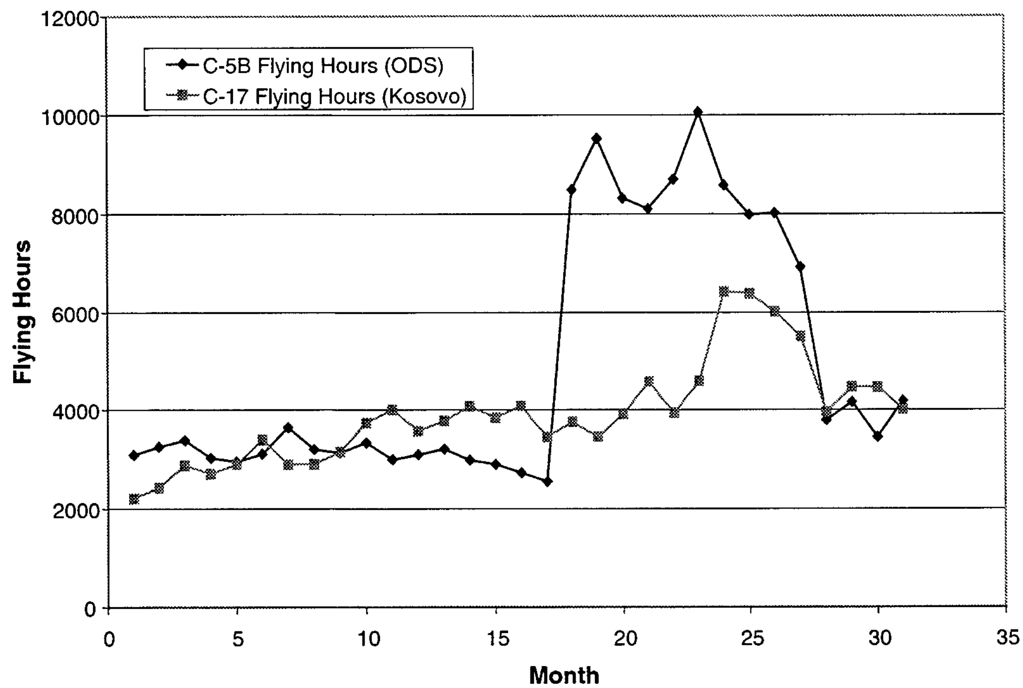
KOSOVO: C–17, KC–135, F–16C

The C–5B data from ODS show an obvious, significant surge that clearly highlights the potential of the physics-based model. Because of the age of the data, however, we had doubts about the data's accuracy. Furthermore, the lack of maintenance personnel with first-hand knowledge of the fleet during ODS makes reaching definitive conclusions about the models difficult with these data alone. We used more recent data from Kosovo to confirm that the physics-based model is more accurate than a proportional model. We chose a transport, a transport/tanker, and a fighter to show that the model applies to a variety of aircraft.

Smaller Surges, Tougher Tests

Before discussing the remaining three aircraft, we must comment on the differences between the ODS data and the Kosovo/Bosnia air operations. The ODS C–5B data show an obvious surge in flying hours and an unmistakable change in flying behavior: Sortie duration increased, and the number of warm take-offs and landings decreased. Obviously, a model that uses flying hours alone is inadequate. Any surges associated with Kosovo are tiny by comparison. Differences in flying behavior also are not as noticeable (Figure 4-6). Surges that do exist are further obscured by the fact that only small portions of a specific fleet of aircraft may have participated in them. These factors make Kosovo data a much tougher test set for the physics-based model than ODS data.

Figure 4-6. Comparison of Operation Desert Storm (C-5B) and Kosovo (C-17) Surges



We obtained flying hour, sortie number, landing, removal, and possession hour data for more than 100 aircraft variants from the REMIS database. We originally planned to use the AFTOC management information system as the source for monthly usage and cost data on depot-level repairables. Because the AFTOC system is relatively new, however, it lacks sufficient component-level historical data in its database for our purposes. As an alternative to AFTOC, we used the REMIS database. REMIS provided 3- to 5-year histories of the monthly on-equipment component removals for all Air Force aircraft. It also provided the other data we needed for the study—flying hours, sorties, landings, and possessed hours.

Of the aircraft we investigated, only a handful showed a meaningful surge. What exactly is a surge? At first, the answer seems obvious: a significant elevation in flying hours over a given time period. Contriving an example for which this criterion alone does not define a surge is easy, however: The number of aircraft could increase in proportion to flight hours, for example. Flight behavior also should change. For aircraft that typically perform touch-and-go as a regular part of training (transports and tankers), the number of warm cycles per sortie should decline during a wartime surge. Other aircraft should see sortie duration rise.

These qualitative guidelines do not translate directly into quantitative rules. We identified significant surges by scaling all four model parameters to flying hours and loosely applying the following guidelines:

- The flying hour-to-ground hour (FH/GH) ratio rises by 30 percent [$(FH/GH)_{scaled} \geq 1.3$].
- The warm cycles-to-cold cycles (WC/CC) ratio falls by 30 percent [$(WC/CC)_{scaled} \leq 0.7$] *or*, if warm cycles are insignificant, sortie duration rises by 30 percent.
- The data set must sustain these values continuously for about 3 months or more.

The first criterion ensures that flying hours show a sustained rise beyond any expected increase because of increased fleet population. The second criterion ensures that the fleet's operational focus has shifted from training and readiness to patrolling and combat. The third criterion ensures that the maintenance activity has time to react to the surge and reach a new steady-state level. With a very short surge, maintenance crews have the option to defer maintenance or cannibalize more, with the possible result that most removals caused by the surge would not take place until after the surge. These criteria may be relaxed somewhat if the surge occurs during a known peacekeeping operation. For example, all of the surges discussed below correspond to *Operation Allied Force/Noble Anvil.*

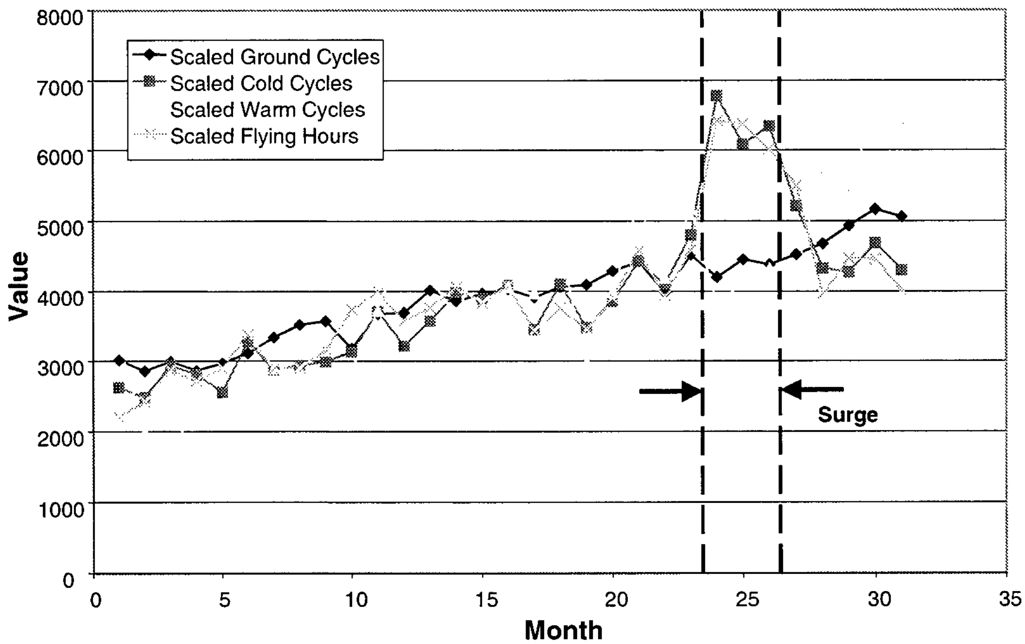
Because the Kosovo operations lack the scale of ODS, very few aircraft fleets showed a significant surge during this period. The C–5B, for example, lacked a surge that meets the three minimal criteria. Its deteriorating reliability record also suggests that the fleet has radically different operational characteristics today than it did 10 years ago. We therefore analyzed the C-17's Kosovo-era data instead. Fighters also posed a problem: No fighters showed a significant surge at the fleet level. Therefore, we narrowed our focus to the F–16Cs stationed at Aviano Air Base in Italy.

Transport: C–17

We requested C–17 data from January 1994 to the present. The REMIS database provided 30 months of reasonably good data for the fleet, beginning with May

1997. Before that, the database provides no removals data. The data for December 1997 lacked removals, so we dropped this point from the analysis. The data show a surge from April through June 1999. These months correspond to months 24–26 in the scaled parameter comparison chart in Figure 4-7.

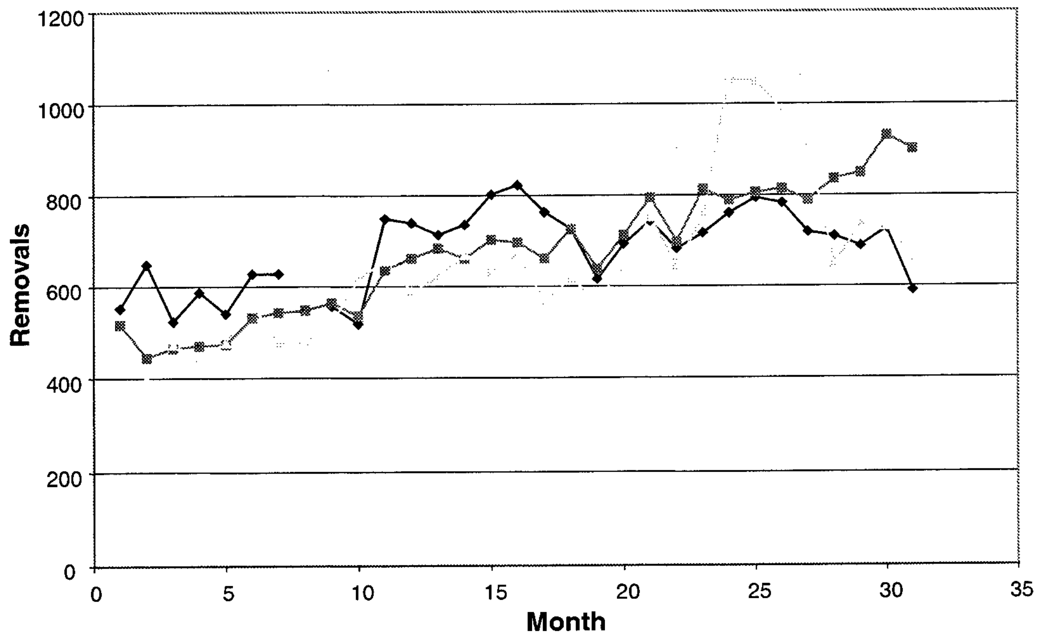
Figure 4-7. Comparison of Model Parameters for C–17 (all parameters scaled to flying hours)



The parameter comparison shows two items of interest: There is strong coupling between flying hours and cold cycles throughout the data set, and warm cycles skyrocket in the four most recent months of the data set. A strong relationship between flying hours and cold cycles is typical of transports and tankers. The second observation is more interesting: What is its cause, and what effect will it have on the model?

The leading-edge calibration set starts at 8 months before the surge and includes all months up to the second month of the surge (months 16–25). The results show that the physics-based model performs slightly better than the proportional model before the surge, much better during the surge, and worse after the surge, when warm cycles are high (Figure 4-8). This apparent disagreement between the model and the data prompted the research team to investigate further.

Figure 4-8. Model Comparison for Kosovo-Era C-17 Data (calibration period: months 16–25)

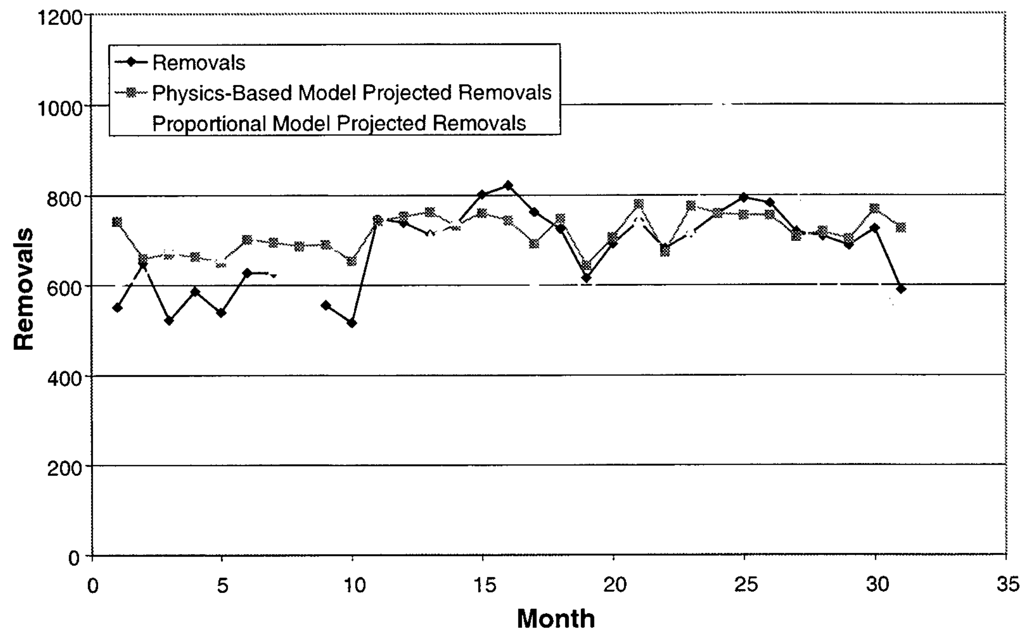


Further investigation shows that the McChord Air Force Base reserve unit is responsible for a disproportionate number of warm cycles in these months. McChord received its first C–17 in July 1999 and had only 5 of 56 C-17s in service by the end of November 1999. During that time, the planes performed more than four warm cycles per sortie; the full fleet performed about 1.7 warm cycles per sortie. The C–17 maintenance chief revealed that the McChord reserve unit is undergoing a rigorous training program to familiarize pilots with the new C–17s. That program is responsible for the dramatic increase in touch-and-gos. If these cycles are less stressful on the planes—perhaps they are not loaded, as is customary for transports during training missions—then the physics-based model would generate artificially high results during those months. The C-17 maintenance chief confirmed that these training missions do not typically involve a full load. (Air Force C–17s typically carry a full load when training.)

To test this hypothesis, we applied the full-fleet model to McChord's C–17s. The model predicted 30 percent more removals than these aircraft actually incurred. When we reduced the probabilities of failure caused by warm and cold cycles by 40 percent, the model tracked reasonably well with the McChord data. The difference is not enough to account for the drop in removals in the later months of the overall data set.

The removals bathtub discussed in Chapter 3 explains the drop in removals. This conclusion is not immediately obvious because overall removals per month seem to rise or remain constant for most of the data set. We expect a young fleet such as the C–17s to show a drop in removals, but the C–17 fleet has expanded continuously for the past several years. If all else remains equal, fleet expansion would make the removals rate rise proportionally. By comparing a trend line of ground cycles data (which closely follows the number of aircraft over time) to a trend line of actual removals data, we can see that actual removals rise significantly slower than expected. The removals bathtub explains this result. We eliminated the bathtub trend by using a similar technique to that outlined in Chapter 3 and compared them to the removals models (Figure 4-9).

Figure 4-9. Model Comparison of Kosovo-Era C-17 Data: Bathtub Trend Removed (calibration period: months 16–25)



As expected, both models fit the data better. The only region of noticeable difference is during the 3 months of the surge: The physics-based model outperforms the proportional model. The difference is not nearly as great as that with the ODS data, but in view of the relative sizes of the surges, that result is not surprising. The physics-based model provides an important advantage in this case: Its three parameters give it the flexibility to react to the data better than the single parameter of the proportional model. For the last 6 months of data, the physics-based model not only fits the data better than the proportional model, it also matches the topology of the model better. Because of this additional flexibility, the physics-based model does not diverge from the data when the fleet undergoes a period of elevated flying hours.

Tanker: KC–10

The REMIS database provided reasonably good data for the requested period of January 1994 to the present—except that October 1994 had a questionably small amount of removals and November and December had no removals. We believe that these three data are errors, and we removed them. The scaled parameter graph shows a surge for the months of April to June 1999 (see Figures 4-10 and 4-11).

The data do not strictly follow our surge definition criteria: The WC/CC ratio is not less than or equal to 0.7 for June. Warm cycles are extremely variable throughout the data set, however, in comparison with the other aircraft in this study. Moreover, the months correspond to *Operation Allied Force/Noble Anvil.* These factors make us feel confident that the physics-based model would show an advantage over the proportional model despite the absence of a surge that meets our criteria.

Figure 4-10. KC-10: All Parameters Scaled to Flying Hours

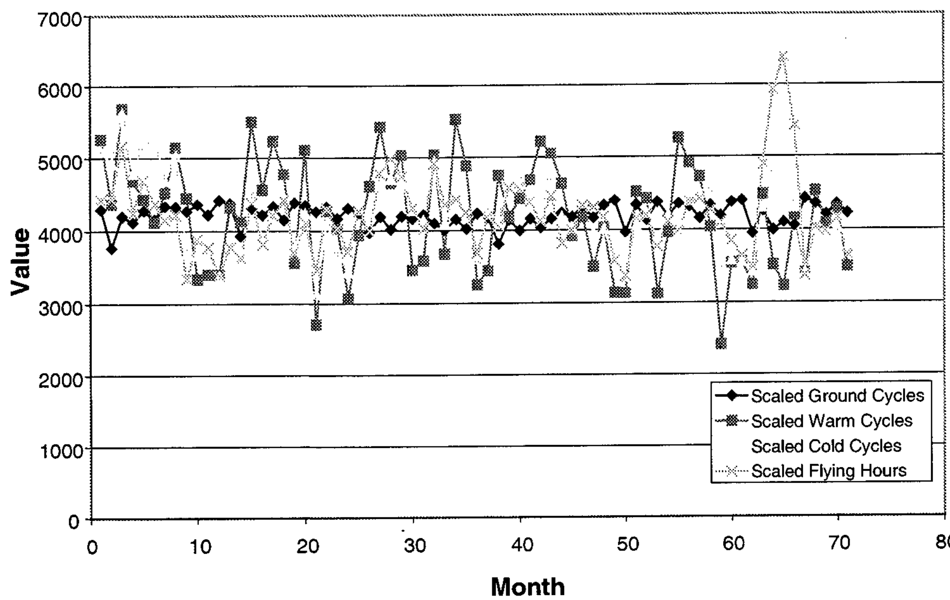
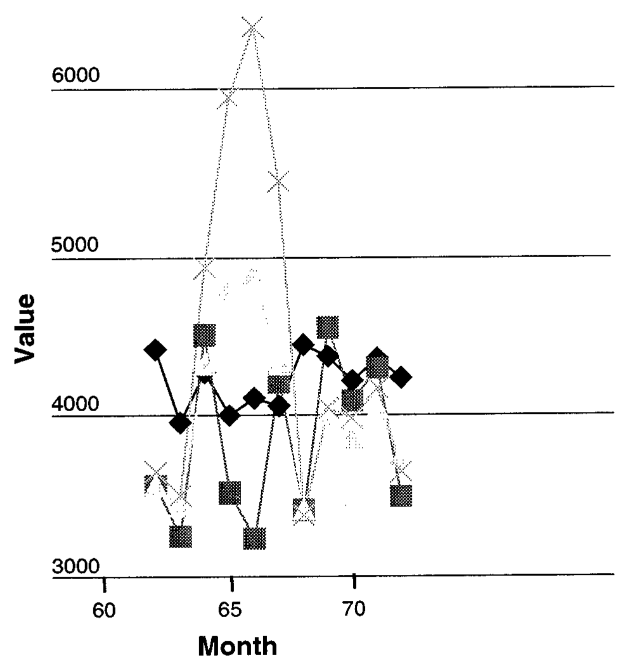


Figure 4-11. KC-10: Operation Allied Force Surge (detail)



As before, cold cycles and flying hours are closely correlated; therefore, we removed flying hours from the model. In accordance with our method of using one third of the data to calibrate the models, we chose to calibrate on months 45–64 for these data. That calibration set includes the first month of the surge and the 19 months prior to that. The physics-based model clearly outperforms the proportional model during the surge period (see Figures 4-12 and 4-13).

Figure 4-12. KC-10: Model Comparison with Actual Removals

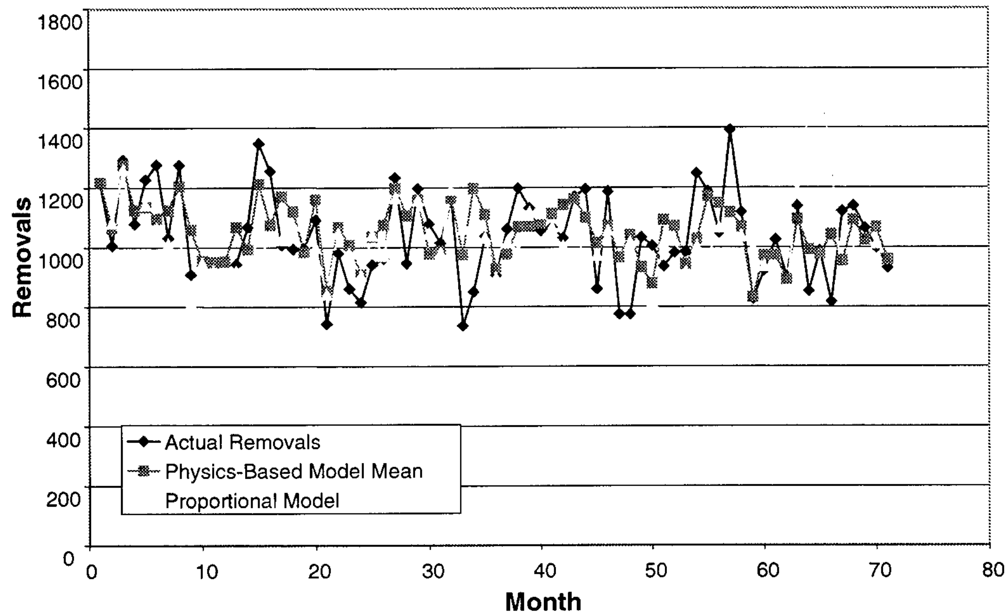
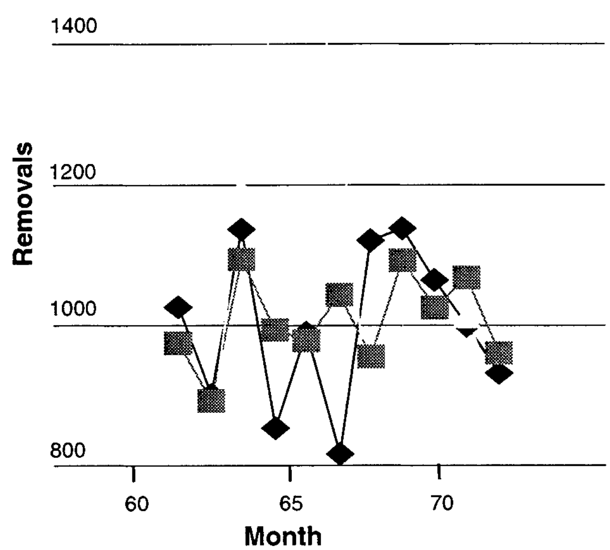


Figure 4-13. KC-10: Model Comparison with Actual Removals (detail)

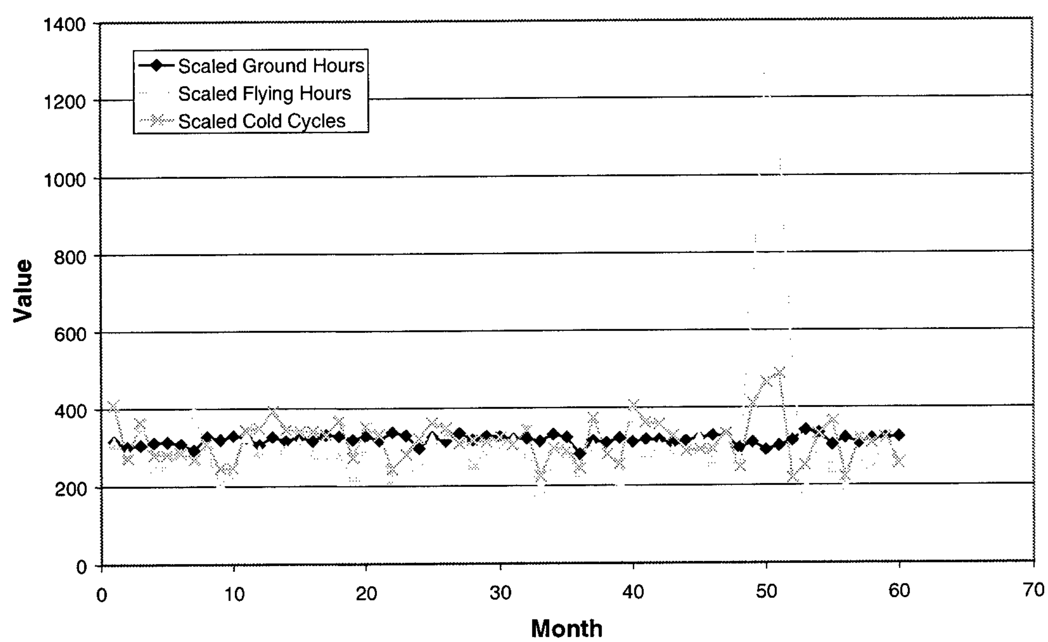


The ground environment accounts for 39.4 percent of total predicted removals; cold cycles account for 20.2 percent, and warm cycles account for 40.4 percent. The physics-based model provides little, if any, improvement except during the surge. The proportional model's mean relative error is 62.5 percent during the surge, and its RMS error is 62.8 percent. These numbers are similar because the model consistently overpredicts actual data during the entire surge. The physics-based model's average relative error is 14.4 percent during this time, and its RMS error is 18.6 percent. These numbers also indicate that this model overpredicts the data, though not nearly to the extent that the proportional model does.

Fighter: F-16C

The F–16C has been a stalwart in the Bosnia/Kosovo air campaigns, but we cannot recognize a surge in the past 6 years of data. The F–16C fleet is so large that only a small portion supported these campaigns at any one time; therefore, no noticeable change in flight behavior exists fleet-wide. F–16Cs assigned to Aviano Air Base have flown many Kosovo missions. Consequently, they show a significant surge during the months of March to June 1999 (Figure 4-14). Unlike the tankers and transports, the F–16C fleet performs practically no touch-and-gos. Therefore, there are too few warm cycles to include in the model. Cold cycles still appear to be highly coupled with flying hours for most of the data set, though some topological differences exist. This is particularly true during the surge. Cold cycles rise for 3 months, whereas flying hours peak, then fall. For this reason, we keep flying hours in the model.

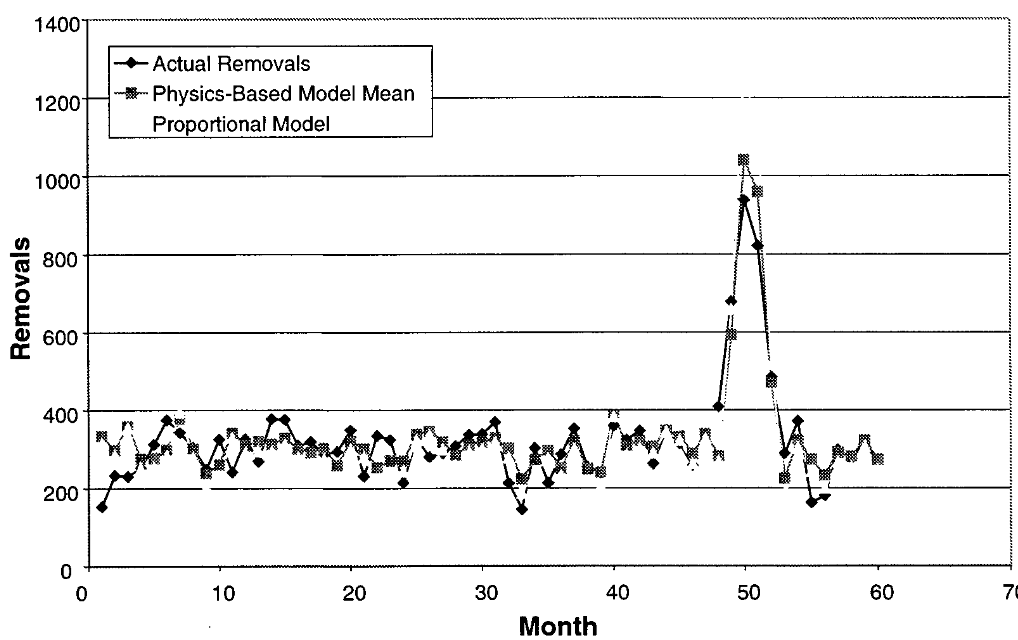
Figure 4-14. F-16C (Aviano): Scaled Parameter Comparison



As with previous data sets, one month (January 1999—month 47 in the set) has missing removals data. We removed this point from the analysis.

In keeping with previous analyses, we selected a calibration range that includes about one-third of the total data and half of the surge. The 20-month calibration range, which starts 18 months before the surge and ends 2 months into the surge, leaves 40 months of data for the test range. The proportional model looks relatively good for this data set (Figure 4-15). The increase in flying hours corresponds to a large increase in removals, so the proportional model has some validity. Without a large number of warm cycles to generate stresses on the plane, the stresses of steady-state flight account for considerably more removals than otherwise. Nevertheless, the mean relative error shows that the proportional model overpredicts actual removals during the surge by 18.8 percent The physics-based model overpredicts the number of removals by 3.1 percent. The proportional model's RMS error is 25.7 percent, whereas the RMS error of the physics-based model is 12.0 percent. From the RMS numbers, we can see that both models capture the shape of the removals curve reasonably well. The proportional model's overprediction of the magnitude of the curve inflates its RMS error to twice that of the physics-based model, however.

Figure 4-15. F-16C (Aviano): Model Comparison with Actual Removals



The physics-based model predicts that flying hours is the dominant parameter affecting removals for Aviano F–16Cs: It accounts for 72.6 percent of removals. Although the ground environment plays somewhat less of a role than it does with the C–17 and the KC–10, it remains a significant source of predicted removals; it accounts for 19.2 percent. Cold cycles account for only 8.3 percent of predicted removals. This result suggests that the lighter fighter undergoes relatively less stress during landings and take-offs than its heavier counterparts.

CALIBRATION RANGE STUDY

The foregoing analysis focuses primarily on leading-edge calibration sets, which include about one-third of the available data and about one-half of the surge data. These data sets give the physics-based model the maximum opportunity to identify the effect of each parameter, and they give the proportional model a chance to average surge data with non-surge data. Therefore, both models usually are more accurate than they would be if the set had no surge data. Analysts who would use these models are not interested, as we are, in calibrating a model with only part of the available data, however. Analysts would use all of the available data to predict removals in upcoming months or years. They may not have the luxury of a noticeable surge in their data. How do the models behave with different calibration sets? To earn our complete confidence, the physics-based model must be able to perform as well as the proportional model regardless of the calibration data. It must be robust.

The two different calibration sets from the C-5B discussion suggest that the physics-based model is robust. To confirm this hypothesis, we divided each aircraft's data set into thirds and calibrated each subset. The results of this exercise appear in "Summary of Results" (below). The Aviano F-16C data sets, because of their relatively high correlation between flying hours and removals, provide the most stringent test of the robustness of the physics-based model; therefore, we discuss the results in detail here. We calibrated on the first 20 months, the middle 19 months (excluding a point with missing removals data), and the last 20 months of the Aviano F–16C data set.

For the first 20 months, the two models perform comparably well. The mean relative error shows that the proportional model overpredicts the 4 months of the surge by 23.5 percent, whereas the physics-based model under-predicts removals by 24.0 percent during the same period. The RMS errors also agree closely; the proportional model's RMS error is 29.7 percent, whereas the physics-based model's is 24.7 percent.

During the second 20 months, the physics-based model is the clear winner. The proportional model overpredicts removals by 25.4 percent, whereas the physics-based model underpredicts removals by 1.8 percent. The proportional model's RMS error is 31.5 percent; the physics-based model's is 10.3 percent. For the third 20 months, the surge lies entirely in the calibration set. Thus, this set provides, for both models, the most accurate approximation of actual surge removals of the three sets. The proportional model still overpredicts removals considerably—by 14.2 percent, in contrast to a 1.2 percent underprediction by the physics-based model. The proportional model's RMS error is 22.1 percent; the physics-based model's is 9.8 percent.

Clearly, the physics-based model does no worse than the proportional model for any of the calibrations sets used. Most often, it does noticeably better. Because flying hours play a significantly more important role with fighters than with transports and tankers, we think that a surge the size of the C–5B's ODS surge would not cause as great a discrepancy between the proportional model and actual removals. Based on the four calibrations we performed, however, we believe that the difference would remain significant—and that the physics-based model would remain better.

AVIANO AND THE FULL F–16C FLEET

The F–16Cs stationed at Aviano Air Base do not show an increasing removals bathtub trend, whereas the entire F–16C fleet does over the same time span. We do not know for certain why this is true. The F–16C/D maintenance chief confirmed that the fleet as a whole has seen a steady rise in removals. Aviano's F–16s may not share this steady rise because of one or more of the following factors:

- They may be younger than the full fleet. We could not confirm whether this is the case.
- They are all Block 40 airplanes. Removals characteristics may vary by block number.
- Aviano is a U.S. Air Force Europe base. As such, its F-16 squadrons receive the highest-priority maintenance. The extra attention may keep those planes newer longer.

The relative number of removals predicted for both populations is essentially the same. The Aviano F–16C model predicts 19.2 percent ground-induced removals, 8.3 percent cold cycle-induced removals, and 72.6 percent flying hour-induced removals. The total F–16C fleet model predicts 20.1 percent ground-induced removals, 6.4 percent cold cycle-induced removals, and 73.4 percent flying hour-induced removals. This similarity in the relative importance of factors affecting removals suggests that the two populations are similar (e.g., one does not have a more reliable engine than the other does) but have different ages.

SUMMARY OF RESULTS

Our study considered ODS data for the C–5B and Kosovo-era data for the C–17, KC–10, and F–16C. We calibrated the physics-based model and the proportional model on four subsets of the data for each aircraft. This procedure provided 16 calibration sets to compare the performance of the two models. For the first three calibrations of each aircraft, we divided the data into three approximately equal subsets. For the fourth calibration, we chose a "leading-edge" subset of data to calibrate. That is, we included (roughly) the first half of the data from a surge, as well as a sufficient number of months directly preceding the surge to provide a calibration set of approximately the same size as the first three sets.

We used two different statistics to measure the accuracy of each model in relation to actual removals: relative error and RMS error. Relative error preserves the sign of the difference between model removals and actual removals; therefore, this statistic measures how well the model follows the central tendency of the data. The RMS error provides a relative measure of the magnitude of the difference between the model and actual removals. Therefore, this statistic measures how well the model follows the actual data at each individual data point.

The results of the trials described above appear in Table 4-1. The physics-based model consistently predicts removals during wartime surges more accurately than the proportional model. The proportional model is sensitive to the size of the surge. When the C–5B experienced a 200–300 percent rise in its flying hours, the difference between predicted and actual removals rose correspondingly. When *Operation Allied Force* prompted a 30–60 percent rise in flying hours in the four planes we studied during that time, the proportional model typically was off by 20–30 percent. The proportional model also showed significant sensitivity to the calibration subset: If the subset includes a part of the surge, the proportional model will better predict removals in the rest of the surge. (See C–5B calibration sets 1 and 4 in Table 4-1.) It will also underpredict removals for all months of normal flying. If the surge is large enough, this underprediction is quite large.

Table 4-1. Predicted Removals by Model Parameter and Comparison of Proportional and Physics-Based Models During Surges

		Predicted removals by model parameter (%)				Proportional model		Physics-based model	
Aircraft	Cal. set	Ground cycles	Cold cycles	Warm cycles	Flying hours	Relative error	RMS relative error	Relative error	RMS relative error
C–5B[a]	1	49.0	25.3	25.7	0.0	203	212	–15.3	22.1
	2	32.8	27.0	40.2	0.0	42.4	51.6[c]	1.0	18.6
	3	30.6	62.7	6.7	0.0	65.9	73.7[c]	14.6	25.2
	4[b]	37.6	25.2	37.2	0.0	94.4	102.3[c]	–4.24	18.38
C–17[a]	1	49.9	39.6	10.5	0.0	24.0	25.5	–2.1	4.5
	2	33.8	40.7	25.5	0.0	38.5	39.4	8.5	9.3
	3	64.0	35.3	.7	0.0	16.6	17.8	.02	.03
	4[b]	53.3	23.4	23.2	0.0	22.6	23.7	–2.4	3.1
KC–10	1	22.8	34.2	43.0	0.0	72.9	73.1	13.6	17.6
	2	36.0	27.4	36.6	0.0	62.0	62.1	11.1	15.5
	3	47.0	17.7	35.3	0.0	62.7	62.9	15.8	19.6
	4[b]	39.4	20.2	40.4	0.0	62.5	62.8	14.4	18.6
F–16C (Aviano AB)	1	53.0	0.0	0.0	47.0	23.5	29.7	–24.0	24.7
	2	29.8	0.0	0.0	70.2	25.4	31.5	–1.8	10.3
	3	32.6	0.0	0.0	67.4	14.2	22.1	–1.2	9.8
	4	19.2	8.3	0.0	72.6	18.8	25.7	3.1	12.0
F–16C[a] (All)	1	20.1	6.4	0.0	73.4	N/A	N/A	N/A	N/A

Note: N/A = not applicable.

[a] "Bathtub" trend identified and removed.

[b] "Leading-edge" calibration set.

[c] Proportional model significantly underpredicted actual data for these calibration sets.

In addition to being more accurate, the physics-based model is more robust than the proportional model. (See C–5B calibration sets 1 and 4 in Table 4-1.) Although it is best to include at least part of a recognizable surge in a calibration set, the model performs well even when the calibration set has no surge data. At worst, the model performs comparably to the proportional model; on average, it performs better.

A closer look at the data reveals that each of the four model parameters may have a significant effect on the outcome, depending on the type of aircraft. It is immediately apparent that the ground environment has a significant effect in all cases, as many maintenance personnel have suspected. Warm cycles are particularly important for transports and tankers. This finding suggests that the key parameter to consider in maintaining these types of aircraft is the number of landings, not the number of flying hours. Warm cycles do not cause significant removals in fighters simply because fighters seldom perform touch-and-gos: One sortie almost always means one take-off and one landing.

Because cold cycles and flying hours most often are closely related, understanding their effect on removals is more difficult. For tankers and transports, we eliminated flying hours from the model because flying hours and cold cycles track too closely to be discernable as separate effects. In view of the limited scope of this investigation, we did not feel that the use of advanced statistical methods to determine their individual effects on removals was warranted. As a result, removals from cold cycles actually indicate the combined effects of cold cycles and flying hours. We know that warm cycles and cold cycles have approximately the same probabilities of failure for these aircraft. We would expect the probability of a cold cycle-induced failure to be much higher if flying hours had a significant effect because flying hour-induced failures are implied in the cold cycle portion of the model. Based on that reasoning, flying hours probably have a more pronounced effect on the KC-10 than on the C-5B and C-17. Further investigation is required to confirm this hypothesis.

Clearly, flying hours are a significant source of failures in the F–16C. Cold cycles—and take-offs and landings in general—do not affect fighters nearly as significantly as they affect the heavier, less maneuverable transports and fighters. In-flight cycles, such as turning on afterburners, may be a significant source of failure-causing removals in these planes. This hypothesis should be investigated in future work.

Chapter 5
Findings and Conclusion

The ability to obtain required performance data is a crucial issue; in general, obtaining data can be the hardest part of any analytical task. Unless analysts can find the data required to calibrate the model, the model isn't feasible. We have found that the REMIS database is an acceptably reliable source for required data. Obtaining the data was relatively easy after we completed the initial administrative process. Air Force analysts could obtain these data even more easily. If analysts want to update the model regularly and periodically, automating the process and reducing the time from a few days to a few hours would be straightforward. In our experience, the database generally offered complete data; a few suspect points were easy to spot, and their removal did not affect the model's calibration.

New data available from AFTOC will be easier for cost analysts to obtain and use. We also expect that this database will provide detailed, complete financial information. With these data available, implementing a model that computes material consumption costs directly from the model's input parameters is feasible.

The model's physics and implementation make it easy to understand and use. It uses two simple algebraic expressions that relate the four inputs to an expected removals value and a variance about that value. The user calibrates the model by performing a maximum likelihood estimation on the data. The user can do this calibration in a matter of minutes with a spreadsheet that has an appropriate optimizer add-on. With additional, one-time effort, automating this process would be possible as well.

We found that many data sets have at least one pair of collinear parameters. In these cases, we have found that eliminating one of the collinear parameters and calibrating the data with the remaining three is adequate. For tankers and transports, we found that eliminating flying hours from the model provided the best results. For fighters, flying hours appear to be a dominant cause of removals, but warm cycles are too few to have significance. Therefore, we modified the physics-based model for tankers and transports to include ground cycles, warm cycles, and cold cycles. For fighters, we modified it to include flying hours, ground cycles, and cold cycles.

Three of the four planes we studied showed a long-term rising or lowering trend in removals. Neither the proportional model nor the physics-based model properly addressed these removals bathtubs. This study showed that removals bathtubs are real, explainable events that can be treated by correcting the data for the trend. Treating them improves the predictive accuracy of each of the models we studied.

In the course of this study, we found that loading (cargo or weapons configuration) may have a significant effect on removals. The C–17s currently stationed at McChord Air Force Base show significantly fewer landing-induced removals than the entire fleet. This finding appears to be attributable, at least in part, to an intense period of training with reduced loads. The C–17 maintenance chief concurred with this conclusion. Furthermore, the F–16 maintenance contact with whom we spoke suggested that the choice of weapons configuration may affect removals.

On the basis of these findings, the most important conclusion we draw from this investigation also is the most obvious: A physics-based model works better than a proportional model. It is more accurate and more robust than the proportional model, and it works on relatively recent data (*Operation Allied Force*) as well as relatively old data (*Operation Desert Storm*).

Chapter 6
Suggestions for Future Work

This report clearly shows that the physics-based model provides a more robust and reliable means of predicting material consumption under all flying conditions. With additional investigation, we should be able to identify new sources of removal-causing failures and understand current sources better. Implementing a spreadsheet model that automates the tasks of gathering required data from the REMIS system and generating projected costs and parts quantities for a given fleet of aircraft also should be useful.

Before we can implement an automated model, we must first fully confirm our assumption that the model's inputs can be accurately forecast. We know that the Air Force currently predicts flying hours with a reasonable degree of confidence; otherwise, implementing flying hour-based proportional models would be impossible. We should investigate not only whether we can predict the other model inputs with a comparable level of confidence but also the difficulties associated with making those predictions.

An automated model also requires an automated means of predicting costs on the basis of removals data that the model generates. In this report, we have established a framework for predicting costs on the basis of removals. In future work, we should implement this framework to determine the difficulties that exist and the level of accuracy we can expect.

The physics-based model is easy to use and understand because it requires only four input parameters. The current model, however, may ignore other events that could be significant sources of removal-causing failures. For example, accelerations above a certain g-force or going to afterburners may cause significant stress cycles on an aircraft. Now that the validity of the basic physics-based model has been verified, investigating whether the inclusion of such data will improve the model or simply add unnecessary complexity should be instructive.

Applying the model in more detail also would be helpful. F–16C data from Aviano Air Base differ significantly from full-fleet data: The Aviano subset does not have a removals bathtub, but the full fleet does. How much variability exists between similar aircraft stationed at different bases? Does the ground environment differ significantly? Are landings noticeably more stressful on aircraft at one base than at another? How does a change in cargo loading or weapons configuration affect the likelihood of removals? Now that we know that the model is valid in general, we can use it to investigate some of these questions.

We also must consider a more detailed investigation into the removals bathtub itself. Is a shear transformation based on a straight line acceptably accurate in all cases, or did it just happen to perform well for the cases we investigated here? What is the true nature of the curve? Is it a set of exponential curves or something else? We also should consider whether we can predict when this phenomenon transitions from a lowering trend to a steady trend and then again from a steady trend to a rising trend.

The logical next step is to create a software application that automates the tasks of obtaining input data, calibrating the model, and investigating the results. LMI has successfully developed several such applications to support other analyses. The conceptual framework of such a tool would be simple; it would have the same form of the spreadsheets and graphs that we already have developed for this project. The analysis spreadsheet would have to be generalized so that it could accept a user-specified calibration set and so data for different aircraft could be swapped into and out of the spreadsheet automatically. This application also would require the capability to automatically upload and preprocess data from its designated source.

Appendix
Calibrating the Model

To calibrate our model, we generate maximum-likelihood estimators of $\lambda_{FH}, P_{GC}, P_{CC}$, and P_{WC}. That is, we choose values for those parameters that maximize the likelihood of an observed sequence of replacements. Thus, given a set of *M* values for replacements *r*, flight hours *FH*, ground cycles *GC*, warm cycles *WC*, and cold cycles *CC*, we choose $\lambda_{FH}, P_{GC}, P_{CC}$, and P_{WC} to solve

$$\max_{\lambda_F, P_{GC}, P_{CC}, P_{WC}} \prod_{i=1}^{M} N(r_i; \mu_i, \sigma_i), \qquad \text{[Eq. A-1]}$$

where

$$\mu_i = CC_i P_{CC} + WC_i P_{WC} + \lambda_F FH_i + GC_i P_C \qquad \text{[Eq. A-2]}$$

and

$$\sigma_i = \sqrt{CC_i P_{CC}(1-P_{CC}) + WC_i P_{WC}(1-P_{WC}) + \lambda_F FH_i + GC_i P_{GC}(1-P_{GC})}\,. \qquad \text{[Eq. A-3]}$$

Because the variances vary from period to period (i.e., they are *intrinsically heteroschedastic*), this maximization does not reduce to a simple method like linear regression. It can be done numerically, however. In fact, it can be done on a simple Excel spreadsheet, using the built-in solver add-on.

To calibrate the model in Excel, the analyst must represent each time interval on a single row of the spreadsheet (see Table A-1). Eq. A-2 calculates the model mean, and Eq. A-3 calculates its standard deviation. With these numbers, the user can use the NORMDIST function in Excel to calculate a probability that the model coefficients are correct given the mean and the actual removals. To compute the most likely set of model coefficients, the analyst must compute the cumulative probability of the entire data set by multiplying the individual probabilities together and varying the coefficients to maximize the result.

Table A-1. Portion of Excel Spreadsheet Used to Calibrate the Model

Actual removals	Model mean	Model variance	Model standard deviation	Probability	Cumulative probability
153	319.6706	310.7993	17.6295	8.834E-22	8.834E-22
234	294.8028	286.357	16.92209	3.70685E-05	3.27463E-26
231	345.4027	336.8391	18.35318	7.94115E-11	2.60043E-36
254	274.6983	265.9526	16.30805	0.010932194	2.84284E-38